THE METHODS OF STATISTICS

AN INTRODUCTION MAINLY FOR EXPERIMENTALISTS

by

L. H. C. TIPPETT, M.Sc.(LOND.)
Statistician to the British Cotton Industry Research Association

Third Revised Edition

LONDON
WILLIAMS AND NORGATE LTD
GREAT RUSSELL STREET

FIRST PUBLISHED IN . 1931
SECOND EDITION, REVISED 1937
THIRD EDITION, REVISED 1941

PRINTED IN GREAT BRITAIN BY
UNWIN BROTHERS LIMITED, LONDON AND WOKING

PREFACE

THE science of statistics in this country seems to have been subject to two main influences. The biometric school associated with Professor Karl Pearson have developed methods and concepts which form the basis of the whole subject, and in more recent years the needs of biological experimentalists have been met by developments in the theory, due largely to Dr. R. A. Fisher. There are many text-books on what may be regarded as the classical theory of statistics, and Fisher's own methods are described in his book, *Statistical Methods for Research Workers*. In the present book I have attempted to present a single system of statistics, so that a reader with little previous acquaintance may obtain a good working knowledge and understanding of the methods available.

The first chapters deal with frequency distributions and constants, and with the theory of errors, in orthodox manner, but in the later chapters the underlying theme is Fisher's idea of the Analysis of Variance; correlation is introduced as a special case of this. There are, of course, other ways of regarding the subject, but its unity seems to me to be brought out more by this than by any other method of presentation.

In outlining the theory underlying the methods described, I have given mathematical proofs where they are easy; but where they are not, I have not hesitated merely to state the results. The necessary mathematical attainments of readers are slight and include algebra up to the binomial theorem and the elementary notions of the calculus; even those who have to omit the mathematics entirely should be able to arrive at an appreciation of the arguments. Nevertheless, the theory of statistics is not easy, not so much because it is abstruse, as because the ideas are new to most people, and a good deal of hard thinking and patient work will be necessary.

I have drawn largely on the literature for examples, and it will be obvious to biologists that I am not one, for some of the statistical deductions I make are of no biological interest, although the methods applied to other inquiries may be of great value. Statistics is a tool, and the good craftsman knows what he wants to make with his tool before he lifts it; it is only the small boy who uses it promiscuously. In this book I have often been like the small boy, but if readers will be small boys with me, they will learn how to handle

the tool, and then will be able to apply it with biological judgı
There is no better way of learning statistics than by working thr
examples.

I wish to thank Mr. F. D. Farrow, Mr. H. A. Hancock
Miss R. Hodgson for their help and criticisms, and Mr. L.
Galloway and Mr. J. Gregory for providing me with unpublis
material for examples.

L. H. C. TIPPE

August, 1931

PREFACE TO SECOND EDITION

FOR this edition I have revised the book considerably, and in particular have re-written the first three chapters. There is now more mathematical theory than before, and although I have not tried to prove everything, I have aimed at giving typical proofs to show in what way the various distributions and methods arise. These proofs are in separate sub-sections, and may be omitted if desired. There are other additions that call for no special comment.

Several topics that were dealt with in the first edition are now excluded. Tests of significance based on the correlation ratio are omitted because they are less convenient than the equivalent z-tests, and the intra-class correlation coefficient because it is now only of historical interest. I have not included the theory of ranking because (1) it is not often needed, (2) when required it is usually for application to small samples whereas the theory given before was for large samples, and (3) I am not aware of any convenient methods involving ranking that are as well founded as the other methods given in this book.

I wish to acknowledge with thanks the criticisms, suggestions and corrections received from many friends and correspondents, and am particularly indebted to Dr. A. J. Turner for reading and commenting on the manuscript of this revised edition.

L. H. C. T.

June, 1937

PREFACE TO THIRD EDITION

THIS year I become aware of new directions in which I think this to t could be improved, and developments in the subject of statistics fes make periodical revisions desirable. The changes I would wish forake at this stage, however, are not sufficiently important to of fy any considerable amount of resetting of the type, particularly the there's a war on," and so this edition is the same as the second bo pt for a few minor corrections—and one important addition. and e addition is a set of tables, given at the end of the book, of *M* normal probability integral, χ^2, t and tables based on Fisher's z to testing the significance of differences between variances. These p es are a shortened and, in some instances, slightly modified form t hose in R. A. Fisher's *Statistical Methods for Research Workers*, d are sufficient for many of the experimenter's everyday require-

ents. I acknowledge with pleasure the generosity of Professor isher and his publishers, Messrs. Oliver & Boyd, in allowing me o reproduce these tables.

L. H. C. T.

March, 1941

CONTENTS

PAGE

PREFACE 7

INTRODUCTION 15

Experimental and Statistical Methods Compared 15
Terminology 16
Populations 16
Theory of Errors 17

CHAPTER I

FREQUENCY DISTRIBUTIONS AND CONSTANTS 19

Frequency Distributions 19
Formation of Frequency Tables 26
Frequency Constants 27
Proportionate Frequencies 28
Position 28
Variability or Dispersion 30
Shape 33
Computation of Moments 34
Moments from Ungrouped Data 34
Moments from Grouped Data 38

CHAPTER II

DISTRIBUTIONS DERIVED FROM THEORY OF PROBABILITY 43

Probability 43
Mathematical Probability 44
Statistical Probability 44
Binomial Distribution 46
Poisson Series 48
Use of the Binomial and Poisson Distributions for Testing Randomness 49
Normal Distribution 54
Determination of Normal Frequencies 56
Relation between Normal Frequencies and Standard Deviation 59
Practical Applicability of the Normal Distribution 60
Non-normal Curves 61
Sampling Distributions 62
Sampling Distribution of Mean 63
Deduction of Sampling Distributions of Mean and Standard Deviation 64

CHAPTER III

PAGE

ERRORS OF RANDOM SAMPLING AND STATISTICAL INFERENCE 67

Random and Representative Samples 67
Tests of Significance 69
Significance of Difference between Two Sample Means 71
General Discussion 74
Groups of Samples 78
Applicability of Normal Sampling Theory 80
Tests of Significance Based on Skew Sampling Distributions 80
Sampling Errors of Various Constants 82
Means of Binomial and Poisson Distributions 82
Measures of Dispersion 84
Constants of Shape 86
Standard Errors of Functions of Statistical Constants 87
Determination of Population Value from Sample 89
Choice of Statistical Constants 91
Method of Maximum Likelihood 95

CHAPTER IV

GOODNESS OF FIT AND CONTINGENCY TABLES 98

Sampling Distribution of χ^2 98
The Additive Nature of χ^2 102
Contingency Tables 104
General Notes on the Distribution of χ^2 10[illegible]

CHAPTER V

SMALL SAMPLES 110

Variance Estimated from Small Samples 110
Significance of Means: The t Test 112
Essential Character of t 114
Significance of Differences Between Means 114
Significance of Differences Between Variances 117
Significance of Variations Between Several Samples 120
Small Samples from the Binomial and Poisson Distributions 121

(iv) *Numbers of Vice-Counties of England and Wales in which* 266 *Species of Plants are found* (*Matthews*, 1923).

Vice-Counties	1–7	–14	–21	–28	–35	–42	–49	–56	–63	–71	Total
Frequency of Species	129	55	21	18	11	10	11	3	7	1	266

(v) *Degree of Cloudiness at Greenwich*, 1890–1904, *for Month of July* (*quoted by Pearse*, 1923).

Cloudiness	0	1	2	3	4	5	6	7	8	9	10	Total
Frequency of Days .	320	129	74	68	45	45	55	65	90	148	676	1 715

(vi) *Number of Rays of Chrysanthemum Leucanthemum* (*Tower*, 1902).

Rays	16	17	18	19	20	21	22	23	24	25	26	27	28
Frequency of Flowers	1	—	—	2	8	17	23	22	21	27	19	16	14

	29	30	31	32	33	34	35	36	37	38	39	Total
	12	10	16	18	29	20	6	6	—	—	2	284

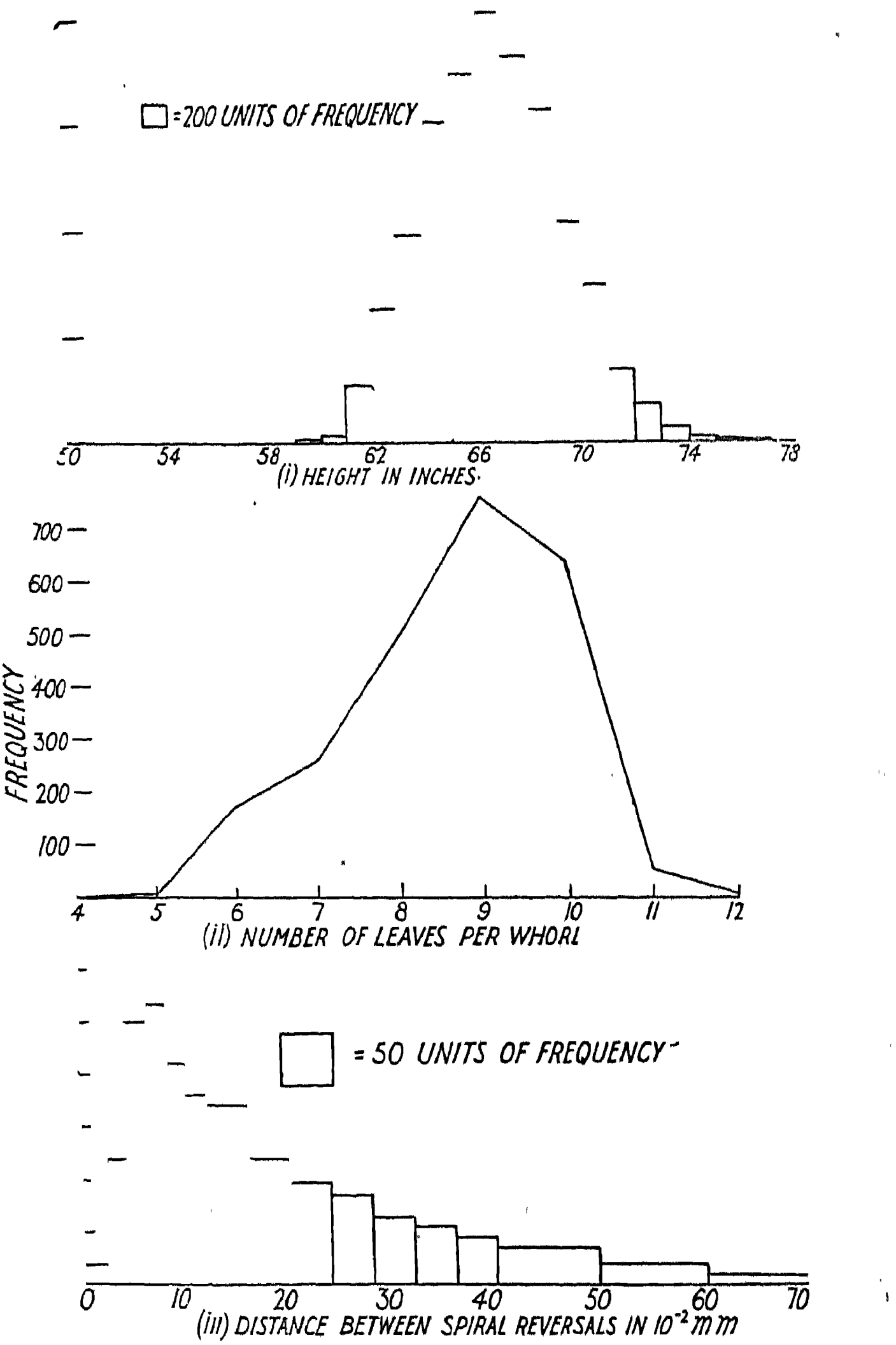
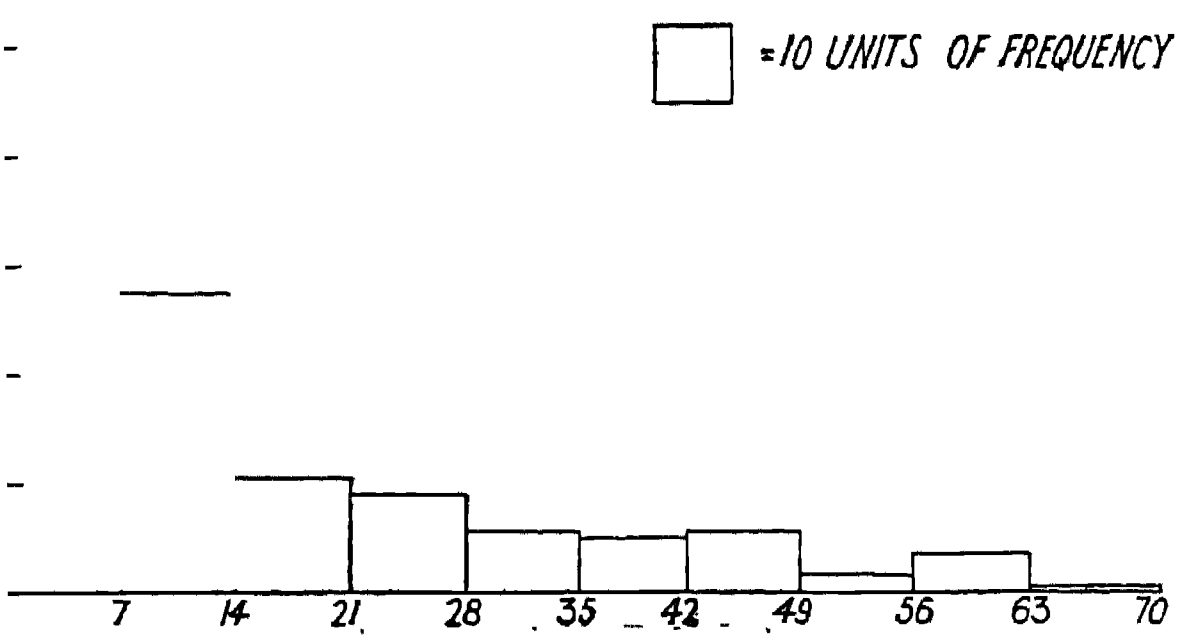
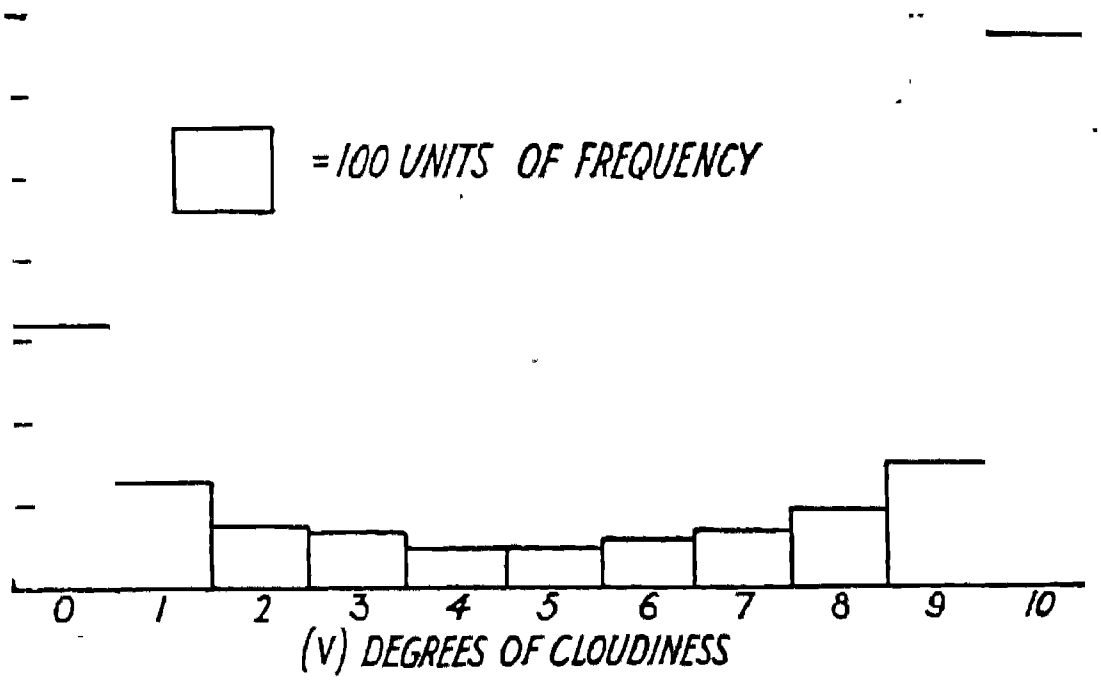


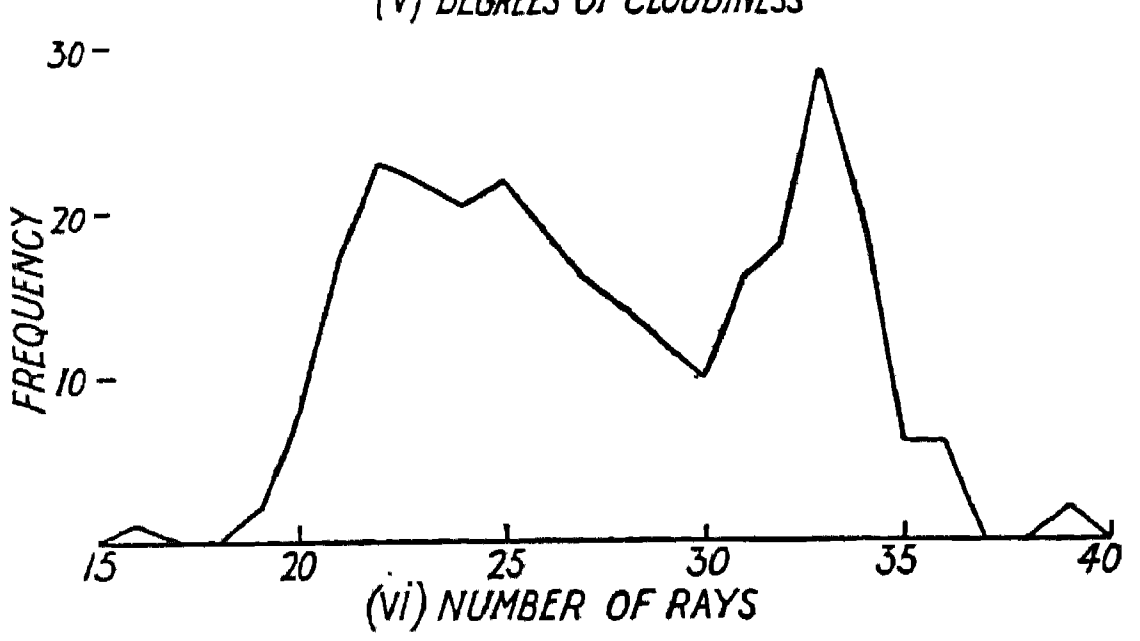
FIG. 1.—FREQUENCY DIAGRAMS.



(iv) NUMBER OF VICE-COUNTIES



(v) DEGREES OF CLOUDINESS



(vi) NUMBER OF RAYS

FIG. 1.—FREQUENCY DIAGRAMS—*continued.*

contains many individuals; indeed for the latter it is the most frequent, showing that most of the 266 species are found only in fewer than eight vice-counties. This is the form of the "hollow curve" mentioned in Willis's book, *Age and Area*, and similar distributions are given by the number of petals of some flowers. The distribution of degrees of cloudiness is of a very unusual form; it shows that for July the sky is usually either nearly completely overcast or completely clear, and that comparatively seldom is there a moderate amount of cloudiness. Sometimes there are two or more humps or *modes* as they are called, as in the rays of chrysanthemums, and such a distribution usually indicates some mixture of types in the data. Care must be taken, of course, not to mistake some small irregularity, such as often occurs, for a second mode.

Formation of Frequency Tables

1.11. The determination of the appropriate size of sub-group of a frequency table or the number of classes into which the total range is divided is a compromise between giving too much and too little detail. If the variate is continuous and the sample is large, it has been found convenient to have between ten and twenty classes, but if the number of readings is less than 100 fewer groups show up better the essential features of the distribution. Care should be taken not to make the grouping too fine relative to the unit of measurement of the character. In dealing with the age returns of men, for instance, it is usually inadvisable to choose sub-ranges of less than a year, for many people give their age last birthday, so that the groups representing integral years contain an undue number of observations, while those representing fractions have comparatively few. If the variate is discrete, the natural unit of grouping is the unit of variation, as in Table 1.1, (ii) and (vi), unless there are too many for the size of the sample and the distribution is irregular; in such instances, several units may be combined to form one group as in the fourth example of Table 1.1, where the 71 vice-counties have been divided into 10 groups.

If the variate is continuous, the sub-ranges must also be continuous; there must be no gap between the highest value in one and the lowest in the next. If this is borne in mind when making the measurements, there need be little difficulty in forming the distribution, but sometimes the data are collected without reference to the fact that they may have to be grouped, and then care is

necessary. A common way of recording data is correct to the nearest unit. Suppose, for example, the data in Table 1.3 are of a continuous variate and are correct to the nearest unit, then the system of classification into groups of three units is unsatisfactory in that there is a gap between the sub-ranges; the first nominally stops at 29 and the second starts at 30 units—where does an individual with an actual value of 29·7 units (say) go? The readings recorded as 29 units may be anywhere between 28·5 and 29·5 units, so the sub-range which nominally extends from 27 to 29 units actually extends from 26·5 to 29·5; the nominal sub-range 30–32 actually extends from 29·5 to 32·5 units, and so on. The actual sub-ranges join up.

Qualitative data may, of course, be classified and formed into a frequency distribution in which the sub-groups are the qualities described; in dealing with hair colour, for instance, red, fair, light brown, dark brown, and jet black may be taken as the separate classes.

Frequency Constants

1.20 Although frequency diagrams are useful as giving a visual impression of the characteristics of a sample, the evidence of the eye alone is not enough, and some numerical measure of the important characteristics should be used for more exact description and comparison. The determination of such a measure carries the process of condensing the original data a stage further than the formation of a frequency distribution, and this can only be done by ignoring further features as being irrelevant to the particular investigation. Sometimes a special constant may be designed for some particular purpose; for example in the kinetic theory of gases, the mean square velocity is a sufficient description of the distribution of velocities of the molecules for the purpose of establishing energy relationships. If the frequency diagram is of unusual shape, e.g. multi-modal, special methods of expression may be necessary, but for most purposes and for distributions commonly met with, there are standard frequency constants that have been devised to express various characteristics. These constants will be described here. It may be emphasised at the outset, however, that in using constants the investigator should always keep the practical problem in view and choose methods that are appropriate. There is no virtue in calculating statistical constants in a blind, routine manner, without knowing first that they will be of some use.

Proportionate Frequencies

1.21. For many purposes, the investigator is only interested in the proportion of individuals having characters below or above a certain value, or between two. It can well be imagined that the U.S. recruiting official would be interested in knowing the proportion of men below (say) 65 inches in height, so that he would know to what extent the number of recruits would fall off if he were to make that the minimum height allowable, and from Table 1.1 (i) we see that the ratio is 6 825 : 25 878, or about 26·4 per cent.

It usually (but not always) happens that the size of the sample is governed by some arbitrary conditions that have no objective significance, so that the distribution is expressed in a more general form if the frequencies are given as fractions or percentages of the total.

When a distribution is represented diagrammatically by a histogram, the proportionate frequency between two limits is the area under the diagram between two ordinates drawn at those limits.

Frequencies and proportionate frequencies underlie nearly all methods of statistical representation. Whatever constants may be calculated or however elaborate may be the analysis, the final interpretation is in terms of frequencies. If the data are sufficiently numerous and the problem is simple enough for an answer in terms of frequencies to be obtained directly, there is no need to compute further constants.

Position

1.22. The location or position of a frequency distribution is some single measure describing a value of the variate about which the observations are scattered. For example, it may be the bull's-eye aimed at by our marksman. There are several constants for describing this.

The most common constant is the *mean* or *average*, which is the sum of all the observations divided by the number.* It is well known and understood, but while admitting the great usefulness of the mean—it is perhaps the most useful constant of all—we would remind readers that it only measures one characteristic of a distribution, and that there are others of great importance.

* This is the arithmetic mean; the geometric and harmonic means are not so much used in statistics.

An alternative measure of position is the median. If all the observations in a sample are ranked in order of value, the median is the value of the middle one if the total number is odd, or a value between the two middle ones if the total number is even. The number of individuals greater than the median value is equal to the number that are smaller, and an ordinate drawn at that value divides a histogram into two equal areas. The median is troublesome to find if the sample is large and the grouping broad, for all the individuals in the group containing the median have to be ranked in order. In the third example of Table 1.1 there are 1 117 observations, so the median is the value of the 559th placed in order, and lies somewhere between 16·5 and 20·5 scale divisions. In that group of 94 individuals the median is the 6th in order of magnitude, and we may take it as being approximately $16 \cdot 5 + \frac{6}{94} \times 4 = 16 \cdot 76$ divisions.

The *mode* is that value of the variate about which the observations are most concentrated, that is the value at which the ordinate of the frequency diagram is highest, and in Table 1.1 (i), for instance, it is between 66 and 67 inches. It is not always easy to define accurately in a sample, for it may be anywhere in the most frequent group, or if the distribution is very flat at the top, and irregular, it is not even easy to decide which would probably be the modal group in the whole population; moreover, a distribution may have more than one real mode, as has already been illustrated. Usually, however, if there is a single mode, the position can be found by assuming Pearson's system of frequency curves as shown in section 1.24.

When the distribution is symmetrical, the mean, median and mode always coincide, and in all other single-modal cases the median comes between the mean and the mode, the dispersion between these three constants being a measure of the extent of asymmetry of the distribution. For practical purposes, the following formula holds approximately if the asymmetry is only moderate:

$$Mean - Mode = 3\,(Mean - Median).$$

Of these constants, the mean is the most fundamental from a theoretical point of view and is the only one that can be used in further analysis of the data. For the mere representation of the central tendency of a distribution, the median is sometimes recommended because it is said to be least affected by extreme individuals, but for most symmetrical uni-modal distributions, the mean is the

most stable constant and is least affected by idiosyncrasies of the particular sample. This result is derived from the theory of errors. When the variate is the life of the individuals under some test of endurance, e.g. the life of electric lamps when burnt at a standard voltage, the median may be an economical constant to determine; for when half the individuals have failed, the median value may be determined without any further testing. Since the modal is the most typical value, it may be the most suitable single constant to use when the distribution is very skew.

Variability or Dispersion

1.23. There are several constants devised to measure the degree of variation or dispersion of the variate, which, in our analogy of section 1.1, is the quality that distinguishes the marksmen. The most fundamental of these are the *second moment*, or *variance*, and its square root, called the *standard deviation*. The second moment is the mean squared deviation from the mean, and may be written

$$\mu_2 = \frac{S(x - \bar{x})^2}{N},$$

where μ_2 is the second moment, x the successive values of the variate in the sample, $\bar{x}$ the mean, N the number of individuals, and S the summation for all values of x. The standard deviation is

$$\sqrt{\mu_2} = \sqrt{\frac{S(x - \bar{x})^2}{N}}$$

and will usually be denoted by the symbol s or σ. When the standard deviation is expressed as a percentage of the mean, it is called the *coefficient of variation*. The computation of some of these constants will be illustrated at the end of the chapter.

Another measure, the *mean deviation*, is often used. It is the sum of the deviations from the mean, irrespective of sign, divided by the number of observations. For frequency distributions of the "normal" form (which will be described in section 2.5), the standard deviation is 1·253 times the mean deviation. This is the basis of Peter's method of estimating the standard deviation.

The *quartile deviation* is the deviation measured on either side of the mean, so chosen that as many observations lie beyond the quartile value as lie between it and the median, and ordinates drawn at the

median and at the two quartiles divide the area under a histogram into four equal parts. If the distribution is asymmetrical, the upper and lower quartile deviations are not equal. The average of the two is half the difference between the two quartile values and is called the *semi-interquartile distance*. In Table 1.1 (iii), if all the observations were placed in order, we should have:

279 obs.	first quartile	279 obs.	median	279 obs.	third quartile	279 obs.

The first quartile lies between 8·5 and 10·5 scale divisions, and the third between 28·5 and 32·5 divisions, and the quartile deviations are the distances of these values from the mean. Historically,

TABLE 1.2

Number in Sample	Mean Range / Standard Deviation
2	1·128
5	2·326
10	3·078
50	4·498
100	5·015
500	6·073

this was one of the earliest measures of dispersion used, but there is no reason why the deviation of the ordinates dividing the frequency diagram in any other proportions should not be used and, indeed, K. Pearson has shown (1920) that the deviation of the ordinate which cuts off $\frac{1}{14}$th of the area determines the dispersion most accurately for the "normal" distribution.

At first sight the *range*, which is the difference between the greatest and smallest members of a sample, would appear to be the most natural index of dispersion; but it unfortunately varies for samples of different sizes taken from the same population, being smaller for the smaller samples. When the population form and sample size are constant, however, the mean range is proportional to the standard deviation, and if their ratio is known the latter can be found from the former. For the "normal" form of distribution,

full tables are given in *Biometrika*, XVII (1925), p. 386, of the ratio

$$\frac{\textit{Mean Range}}{\textit{Standard Deviation}}$$

for samples of between 2 and 1 000 individuals, and an abstract of them is given in Table 1.2; from this we may see how much the range depends on the size of the sample. Of course, single values of the range are liable to rather large chance errors, but if it is convenient

TABLE 1.3

	54	55	41	70	42	50	53	31	49	47
	49	59	51	56	40	52	52	56	41	28
	50	49	57	57	44	53	72	56	53	60
	53	54	32	47	41	43	56	59	53	47
	70	62	56	44	54	49	50	35	53	38
Means .	55·2	55·8	47·4	54·8	44·2	49·4	56·6	47·4	49·8	44·0
Ranges .	21	13	25	26	14	10	22	28	12	32
	54	27	39	52	49	57	28	48	36	48
	27	46	58	65	49	38	51	42	48	47
	48	60	66	60	44	55	49	46	55	42
	41	62	59	41	49	56	50	62	53	69
	47	48	63	62	64	47	53	59	47	46
Means .	43·4	48·6	57·0	56·0	51·0	50·6	46·2	51·4	47·8	50·4
Ranges .	27	35	27	24	20	19	25	20	19	27

to collect the data in groups of 5 or 10 (say), the mean range can be found quite quickly, and when divided by the appropriate constant obtained from the tables, can be converted to the standard deviation. The ratio does not seem to be very sensitive to moderate changes in the form of the distribution, and this method may be used in most practical experience when the frequency distribution is uni-modal and tails off fairly gradually to zero at the extremes. Because of its ease of computation, the range is convenient in routine work. In Table 1.3 there are 100 individuals from an artificially constructed "normal" population, and these are arranged in random groups of 5. The ranges are given, and their mean is 22·3. The constant given in Table 1.2 for samples of 5 is 2·326, whence we obtain for an estimate

of the standard deviation a value $22{\cdot}3/2{\cdot}326 = 9{\cdot}6$; this is not far from the true standard deviation, which happens to be 10 units.

This method of obtaining the standard deviation is only equivalent to the other method when the individuals are mixed and divided into sub-samples at random.

All these measures of dispersion may be rather bewildering to the reader, but they are simply so many alternatives, which are exactly related for any given form of frequency distribution. In view of the theory of the next chapter and section 3.7, it will be well to regard the standard deviation as the fundamental measure of scatter, and the other constants as approximations to it. It is sometimes thought that these measures of variability apply only to distributions of the "normal" type, but this is not so; they are equally applicable to skew distributions, and, provided the shape is the same, may be used to compare one distribution with another.

A qualitative appreciation of dispersion or variation may be acquired with little experience, and for comparative purposes any of the above measures may be appreciated fairly easily; the more precise interpretation of the standard deviation in terms of frequencies is dealt with in section 2.52.

Shape

1.24. The shape of a uni-modal frequency distribution may vary in two ways, in the degree of asymmetry, or in the flatness of the mode. This flatness of the mode (or kurtosis) is different from that flatness of the curve as a whole which arises from the dispersion, and is illustrated in Fig. 2, where there are several curves having the same standard deviation but varying kurtosis. These properties may be measured by constants derived from the third and fourth moments of the distribution. The third moment,

$$\mu_3 = \frac{S(x - \bar{x})^3}{N}$$

and the fourth moment,

$$\mu_4 = \frac{S(x - \bar{x})^4}{N},$$

where the symbols on the right-hand side of the equations have the same meaning as those used in defining the second moment. K. Pearson has derived two constants from the moments, which are inde-

pendent of the dispersion of the distribution, and describe its shape. They are

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2};$$

β_1 is zero if the distribution is symmetrical, while β_2 is usually about 3 for a curve like that of Fig. 1.1 (i), is smaller if the mode is flatter, and larger if the curve is more sharply peaked. In Fig. 2 are given a few smooth frequency curves with different values of β_1 and β_2. They all have positive skewness, and a similar series with negative skewness can be imagined.

The degree of skewness may be more simply measured by the quantity

$$Skewness = \frac{Mean - Mode}{Standard\ Deviation}.$$

This quantity is negative if the curve shows a bias to the right (assuming that ascending values of the variate are taken from left to right) and has the longer tail to the left. The position of the mode, as we have shown, is not easy to define, but if the curve comes within Pearson's* system, the following formula may be used:

$$Skewness = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

in which $\sqrt{\beta_1}$ is given the same sign as μ_3. From this we obtain the equation for fixing the mode of a distribution relative to its mean,

$$Mean - Mode = \frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

These constants of shape are rather difficult to interpret, and it is only after considerable experience that their practical import can be appreciated.

Computation of Moments

Moments from Ungrouped Data

1.31. It is always possible to compute the mean and the higher moments in a sample by straightforward evaluation of the expressions given in sections 1.22–1.24 as definitions. For example, the

* This will be mentioned in section 2.6.

mean and second moment are calculated directly in the first three columns of Table 1.4 for a small sample of ten individuals. We have

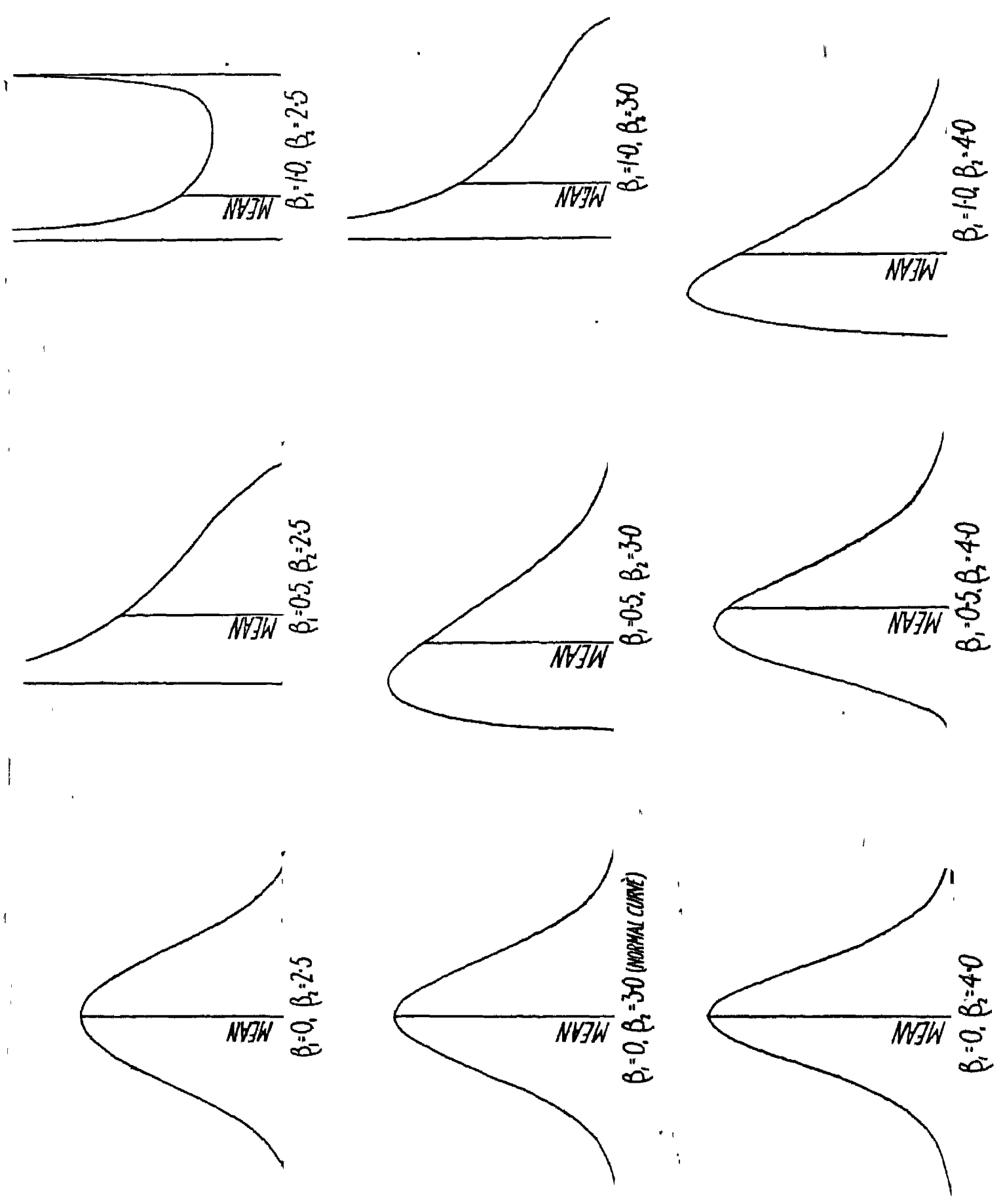


FIG. 2.—FREQUENCY CURVES.

not yet dealt with the theory of methods for such small samples, and this one is given here merely to illustrate arithmetical processes that are applicable to small and large samples. The sums of the

columns are given at the foot of the table, and from them it is seen that the mean, $\bar{x} = 71 \cdot 8$ and the second moment $\mu_2 = 12 \cdot 96$.

Such calculations are much facilitated by making a transformation of the values of the variate, measuring them as deviations from some arbitrary origin or "working mean" and dividing the deviations by an arbitrary constant; this last step is equivalent to representing

TABLE 1.4

x	$(x - \bar{x})$	$(x - \bar{x})^2$	x'	x'^2
67	− 4·8	23·04	− 1	1
70	− 1·8	3·24	0	0
76	+ 4·2	17·64	+ 2	4
73	+ 1·2	1·44	+ 1	1
67	− 4·8	23·04	− 1	1
73	+ 1·2	1·44	+ 1	1
70	− 1·8	3·24	0	0
73	+ 1·2	1·44	+ 1	1
70	− 1·8	3·24	0	0
79	+ 7·2	51·84	+ 3	9
718	0·0	129·60	+ 6	18

the variation on an arbitrary scale. If x' is the transformed variate, X the arbitrary origin, and h the arbitrary constant or scale,

$$x' = \frac{x - X}{h} \quad \text{and} \quad x = X + hx' \quad . \quad . \quad . \quad (1.1)$$

Further, let the mean of the values of x' be $\bar{x}'$, and the mean of their sth power be μ'_s, so that

$$\bar{x}' = \frac{Sx'}{N} \quad \text{and} \quad \mu'_s = \frac{Sx'^s}{N}.*$$

* The letter μ means that the constant is the mean of some power of deviations, the subscript s denotes the order of the power (e.g. for the second moment, $s = 2$) and the prime indicates that the deviations are from an arbitrary origin in arbitrary units. Consistent with this notation $\bar{x}' = \mu'_1$. Where there is no prime, the deviations are measured from the mean in the units of the variate, and the μ_s is the sth moment.

Then it may be shown that:

$$\begin{aligned} \bar{x} &= X + h\bar{x}', \\ \mu_2 &= h^2(\mu_2' - \mu_1'^2), \\ \mu_3 &= h^3(\mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3), \\ \mu_4 &= h^4(\mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4). \end{aligned} \qquad \ldots \quad (1.2)$$

To illustrate the use of these equations the mean and second moment have been calculated for the data of Table 1.4. The arbitrary origin $X = 70$, the constant $h = 3$, and the values of x' are given in the fourth column. From their sum, $\bar{x}' = 0{\cdot}6$ and

$$\bar{x} = 70 + 3 \times 0{\cdot}6 = 71{\cdot}8.$$

The sum of the values of x'^2, given in the fifth column, gives $\mu_2' = 1{\cdot}8$, and from equation (1.2),

$$\mu_2 = 9(1{\cdot}8 - 0{\cdot}36) = 12{\cdot}96.$$

These results agree with those calculated directly, as indeed they should. Readers are recommended to calculate the third and fourth moments for Table 1.4 by the two methods.

1.311. The proof of equations (1.2) depends on applying the ordinary rules of addition to summations denoted by the sign S.* The first is that the summation of a compound term itself made up of the sum of several terms is equal to the sum of the summations of the separate terms. For example,

$$S(X + hx') = SX + Shx'.$$

The second rule is that the summation of a term that is multiplied by a factor constant for all values of the term summed, is equal to that constant multiplied by the summation of the term.

For example,

$$Shx' = hSx'.$$

The general significance of these rules should be appreciated.

From equation (1.1) and the above examples of the summation

* Readers will do well to master these rules for the sake of following later chapters.

rules the first part of equation (1.2) follows easily. It also follows that

$$x - \bar{x} = h(x' - \bar{x}'),$$

and

$$\mu_s = \frac{S(x - \bar{x})^s}{N} = \frac{Sh^s(x' - \bar{x}')^s}{N}.$$

Taking the constant term outside the S sign and expanding the term in brackets by the binomial theorem, we find that

$$\mu_s = \frac{h^s}{N} S\left\{x'^s - sx'^{s-1}\bar{x}' + \frac{s(s-1)}{2!}x'^{s-2}\bar{x}'^2 + \ldots (-1)^{s-1}sx'\bar{x}'^{s-1} + (-1)^s\bar{x}'^s\right\}.$$

Summing this term by term, and remembering that $\bar{x}' = \mu_1'$, and is constant for all individuals in the sample, we combine the last two terms, and find that

$$\mu_s = h^s\left\{\mu_s' - s\mu_{s-1}'\mu_1' + \frac{s(s-1)}{2!}\mu_{s-2}'\mu_1'^2 + \ldots (-1)^{s-1}(s-1)\mu_1'^s\right\}.$$

If s is successively put equal to 2, 3 and 4, the last three of equations (1.2) result.

Moments from Grouped Data

1.32. When calculating the moments of a frequency distribution, it is usual to assume that all the individuals in any one group have the central value of that group, i.e. the value midway between the limits of the sub-range. For Table 1.1 (i) the central values are 51·5, 52·5, etc., inches. These values may be transformed as shown in section 1.31. Then to calculate the moments, instead of summing over all individuals, it is possible to multiply the appropriate power of the transformed group value by the frequency and to sum these products over all groups. Let x_t' be the tth group value (the subscript denotes that the transformed variate can only take discrete values corresponding to the central values of the groups), n_t the frequency in that group, and $\underset{t}{S}$ be the summation over all groups, then for such data,

$$Sn_t = N \quad \text{and} \quad \nu_s' = \frac{Sx'^s}{N} = \frac{Sn_tx_t'^s}{N} \quad . \quad . \quad . \quad (1.3)$$

The letter ν' is used for the mean of a power of a transformed variate calculated from grouped data where μ' would be used if the data were ungrouped.

For example, the data in Table 1.4 may be grouped and the sum of the fourth column is

$$2 \times (-1) + 3 \times 0 + 3 \times 1 + 1 \times 2 + 1 \times 3.$$

Using equations (1.3) and applying equations (1.2) to the resulting means as shown above, it is possible to compute the crude mean and moments from grouped data. These moments are denoted by ν_2, ν_3 and ν_4. For a continuous variate some of the crude moments ν differ slightly from the true moments denoted by μ, even when calculated from very large (infinite) samples, because the discontinuous form of a frequency distribution with finite groups is only an approximation to the continuous form that is possible with a continuous variate. Sheppard's corrections, when applied to such crude moments calculated from a table with uniform sub-ranges, give values that approximate more closely to the true moments. The corrected mean and moments so obtained are:

$$\begin{aligned} \bar{x} &= \textit{crude mean}, \\ \mu_2 &= \nu_2 - \tfrac{1}{12}h^2, \\ \mu_3 &= \nu_3, \end{aligned} \qquad \ldots \quad (1.4)$$

and

$$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_2 h^2 + \tfrac{7}{240}h^4.$$

The above transformations make the computation of moments from grouped data a comparatively easy matter if the arbitrary origin X is chosen to be the central value of one of the groups near the centre of the whole distribution and h is made equal to the sub-ranges, so that the values of the transformed variate x_i' are 0, 1, 2, 3 . . . -1, -2, -3 . . . etc. The corrections (1.4) can then be applied to the arbitrary moments, putting h equal to 1, and the final true moments be found by multiplying by h, h^2, h^3 and h^4; the β coefficients, being ratios, are found directly.

The whole process is followed in the example of Table 1.5, the data being the heights of 1 078 fathers (Pearson and Lee, 1903). Column (1) contains the sub-ranges, and column (2) the values of the centres of the sub-ranges in the arbitrary units, x_i'. Column (3) contains the frequencies (the 0·5 frequencies arise because when an observation falls exactly on the border-line between two groups,

TABLE 1.5

(1) Stature in Inches	(2) Arbitrary Units x'_t	(3) Frequency n_t	(4) $n_t x'_t$	(5) $n_t x'^2_t$	(6) $n_t x'^3_t$
58·5–	— 9	3	— 27	243	— 2 187
59·5–	— 8	3·5	— 28	224	— 1 792
60·5–	— 7	8	— 56	392	— 2 744
61·5–	— 6	17	— 102	612	— 3 672
62·5–	— 5	33·5	— 167·5	837·5	— 4 187·5
63·5–	— 4	61·5	— 246	984	— 3 936
64·5–	— 3	95·5	— 286·5	859·5	— 2 578·5
65·5–	— 2	142	— 284	568	— 1 136
66·5–	— 1	137·5	— 137·5	137·5	— 137·5
67·5–	0	154	—	—	—
68·5–	1	141·5	141·5	141·5	141·5
69·5–	2	116	232	464	928
70·5–	3	78	234	702	2 106
71·5–	4	49	196	784	3 136
72·5–	5	28·5	142·5	712·5	3 562·5
73·5–	6	4	24	144	864
74·5–	7	5·5	38·5	269·5	1 886·5
Total	—	1 078	— 326	8 075	— 9 746

$$\nu'_1 = -326/1\,078 = -0{\cdot}302\,41$$

$$\begin{aligned} \nu'_2 = 8\,075/1\,078 &= 7{\cdot}490\,72 \\ -\nu'^2_1 &= -0{\cdot}691\,45 \\ \hline \nu_2 &= 7{\cdot}399\,27 \\ &= -0{\cdot}083\,33 \\ \hline \mu_2 &= 7{\cdot}315\,94 \end{aligned}$$

$$\sigma = 2{\cdot}704\,8 \text{ inches}$$

$$\begin{aligned} \nu'_3 = -9\,746/1\,078 &= -9{\cdot}040\,8 \\ -3\nu'_2\nu'_1 &= +6{\cdot}795\,8 \\ \hline &= -2{\cdot}245\,0 \\ +2\nu'^3_1 &= -0{\cdot}055\,3 \\ \hline \nu_3 = \mu_3 &= -2{\cdot}300\,3 \end{aligned}$$

$$\beta_1 = (-)0{\cdot}013\,513$$

TABLE 1.5—*continued*

(7) $n_t x_t'^4$	(8) $w=\frac{x_t-\bar{x}}{\sigma}$	(9) A_w	(10) NA_w	(11) n_t (expected)	(12) z	(13) $y=z\frac{N}{\sigma}$
19 683	−3·030 8	0·001 22	1·3	1·3	0·004 0	1·6
14 336	−2·661 0	0·003 90	4·2	2·9	0·011 6	4·6
19 208	−2·291 3	0·010 97	11·8	7·6	0·028 9	11·5
22 032	−1·921 6	0·027 33	29·5	17·7	0·063 0	25·1
20 937·5	−1·551 9	0·060 34	65·0	35·5	0·119 7	47·7
15 744	−1·182.2	0·118 56	127·8	62·8	0·198 3	79·0
7 735·5	−0·812 5	0·208 25	224·5	96·7	0·286 8	114·3
2 272	−0·442 8	0·328 96	354·6	130·1	0·361 7	144·2
137·5	−0·073 1	0·470 86	507·6	153·0	0·397 9	158·6
—	0·296 7	0·616 65	664·7	157·1	0·381 8	152·2
141·5	0·666 4	0·747 42	805·7	141·0	0·319 5	127·3
1 856	1·036 1	0·849 92	916·2	110·5	0·233 2	92·9
6 318	1·405 8	0·920 10	991·9	75·7	0·148 5	59·2
12 544	1·775 5	0·962 09	1 037·1	45·2	0·082 5	32·9
17 812·5	2·145 2	0·984 03	1 060·8	23·7	0·040 0	15·9
5 184	2·514 9	0·994 04	1 071·6	10·8	0·016 9	6·7
13 205·5	2·884 6	0·998 04	1 075·9	6·4	0·006 2	2·5
179 147	—	—	—	1 078·0	—	—

$$\nu_4' = 179\,147/1\,078 = 166\cdot185$$
$$-4\nu_3'\nu_1' = -\ 10\cdot936$$
$$155\cdot249$$
$$+6\,\nu_2'\nu_1'^2 = +\ 4\cdot110$$
$$159\cdot359$$
$$-3\,\nu_1'^4 = -\ 0\cdot025$$
$$\nu_4 = 159\cdot334$$
$$-\tfrac{1}{2}\nu_2 = -\ 3\cdot700$$
$$155\cdot634$$
$$+\ 0\cdot029$$
$$\mu_4 = 155\cdot663$$
$$\beta_2 = 2\cdot908\,3$$
$$\text{mean} - \text{mode} = -\ 0\cdot170\,1$$
$$\text{mean} = 68\cdot0 - 0\cdot302\,4 = 67\cdot697\,6 \text{ inches}$$
$$\text{mode} = 67\cdot867\,7 \text{ inches}$$

a half is put into each), and columns (4), (5), (6) and (7) are obtained successively from the previous one by multiplying by the corresponding units in column (2). In column (4) $-27 = 3 \times -9$, in column (5) $243 = -27 \times -9$, in column (6) $-2\,187 = 243 \times -9$ and in column (7) $19\,683 = -2\,187 \times -9$, and so on for the other rows. The arithmetic may be checked by multiplying each frequency directly by $x_t'^4$, and so checking column (7) term by term; if that is correct the other columns from which that was obtained must almost certainly be right. The sums of these columns (paying regard to sign) give the ν' coefficients, and these are corrected step by step below the table. The resulting moments are in inch-units already, since the sub-groups are 1 inch wide. The mean — mode is found from the formula on p. 34; $\nu_1' = -0{\cdot}302\,4$, and so the mean is $0{\cdot}302\,4$ inch less than the arbitrary origin, which is $68{\cdot}0$ inches (the centre of the group at which $x_t' = 0$). Hence *mean height* $= 68{\cdot}0 - 0{\cdot}302\,4 = 67{\cdot}697\,6$ inches. We shall refer to columns (8) to (13) of Table 1.5 in the next chapter.

In this example, the constants have been calculated correct to several decimal places. This has been done advisedly, for when complicated computations are performed, errors due to "dropping figures" too soon are apt to accumulate and become very large, and it is always well to be on the safe side. The number of figures used in computing this example is near the minimum advisable. The final result may be "rounded off" to a few figures, if it is not going to be used in further calculations.

CHAPTER II

DISTRIBUTIONS DERIVED FROM THEORY OF PROBABILITY

Probability

2.1. There are events about which our knowledge is so complete that we are able to predict with complete certainty whether or not they will occur, and on the other hand there are events about which we know nothing, so that we are unable to make any prediction. An event of the first kind is the falling of a penny tossed into the air—we are certain it will fall—and one of the second kind is the falling of the penny with the head (say) uppermost. Between these two extremes there are events about which we know something, but not enough to allow of certain prediction; these are the province of the theory of probability. The extent to which we can rely on a prediction varies in degree for different events depending on the amount of knowledge we have of the factors that determine the event, and a measure of the degree is called the probability of the event.

Probability is expressed on a somewhat arbitrary scale of numbers between unity and zero, a value of 1 or 0 corresponding to certainty that the event will or will not occur, and intermediate values to intermediate degrees of certainty; a probability of 0·5 means that the event is as likely to occur as not. Defined in this vague way, there seems to be little objectivity about probability as a measure, but precision is achieved by giving the word a special meaning and concreteness by applying it to a restricted field.* These two steps result in mathematical and statistical views of probability. By extending the conceptions of these special fields to the events of ordinary life, the probability of any event is usually expressed as chances for or against the event, and the analogy makes the idea of probability fairly definite and concrete. We shall see how this arises in the next two sections.

* A general calculus of probabilities has been developed for application to the general problem of expressing the relations between events and the amount of knowledge of the determining factors, or between propositions and the data on which they are based. The application of this calculus has not received universal acceptance, but this need not concern us, as we are only interested here in the mathematical and statistical applications that are almost, if not quite, universally accepted.

Mathematical Probability

2.11. This view of probability is based on a conception of a number of equally likely chances, some of which are favourable to the occurrence of the event and some of which are unfavourable. The ratio of the favourable to the total chances is the probability of the event. This leads to a similar scale to that mentioned above, varying between zero and unity. It is usual and helpful to imagine an experiment like that of throwing a perfect six-sided die, all sides of which are equally likely to turn up. Only one side has a six, so the probability of turning up a six is $\frac{1}{6}$. Other idealised games of chance suggest other typical experiments:—drawing blindfold from a bag of well-mixed balls that differ only in colour, or drawing cards from a well-shuffled pack, or spinning a perfectly balanced roulette wheel.

The calculus of mathematical probabilities is purely concerned with finding the probability of composite events knowing those of the simple components, and as such it is an exercise in permutations and combinations, enumerating the chances for and against the composite event. There are two fundamental rules that should be understood.

Rule I.—The probability of occurrence of one or other of a number of *independent* events, only one of which can occur at a time, is the sum of the probabilities of the separate events. For example, it is easy to see that when throwing a die, two of the six equally likely chances favour the turning up of either a five or a six and that the probability is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Rule II.—The probability of the simultaneous occurrence of a number of *independent* events is the product of their separate probabilities. Thus, when throwing two dice, there are 36 equally likely possibilities, and of these, only one is favourable to a double six; the probability of a double six is $\frac{1}{36}$, and this is $\frac{1}{6} \times \frac{1}{6}$.

The word *independent* has its ordinary English meaning in the above rules.

Sometimes the chances for and against an event are specified instead of the probability; a probability of $m/(m+n)$ may be described by the statement "the chances are m to n for the event."

Statistical Probability

2.12. The mathematical laws of probability find much use in the theory of statistics for calculating the relations between samples and

populations. Superficially, and from the statistical standpoint the drawing of a sample of (say) 1 000 men from the population of England is very like drawing balls from a bag, and it is assumed that the technique of sampling may be made to satisfy the essential conditions for the application of the laws of probability. Then, the occurrence of an event becomes the drawing of an individual of a given character, the set of equally likely chances becomes a population of equally likely individuals, and a given combination of events becomes a sample of given composition. Samples which follow the laws of mathematical probability are called *random samples*; they satisfy the essential condition that every individual must be independent; i.e. its character must not be influenced in any way by the character of any other individual in the sample.

On this, the statistical view, the probability of drawing an individual of a given character is the proportion of individuals in the population having that character; statistical probability is a proportionate frequency. Conversely, any probability statement may be interpreted in those terms. When we say that the probability of a given horse winning a given race is one-fifth, we may imagine a very large number or infinite population of races run under the same conditions, in one-fifth of which races the given horse wins.

This interpretation of probability is a little more concrete and so is a little more satisfying practically than its most general description as a ratio expressing a degree of confidence or knowledge, or a ratio of chances, but it is still not quite satisfactory. A population is rarely known, even when it is such a concrete thing as a bulk of corn that is being sampled, for if it were known, there would be no question of sampling.* However, it has been found in statistical experience that small samples from the same bulk vary among themselves considerably, but that as they increase in size they become more stable and are less subject to sporadic variations. It is assumed that this tendency continues indefinitely as the size of the sample increases, that there is a single limiting form to which a sample from a given population tends as the sample increases in size. This limiting sample is what the statistician conceives of as the *infinite population*, and in more ordinary language may be described as the result that would be obtained in the "long-run" of experience. The probability of a given kind of individual is the proportion of

* We are only dealing with instances in which the population is very large compared with any samples that are likely to be drawn.

individuals of that kind in this infinite population. We do expect, in the long-run, that events having given probabilities will occur and fail in nearly the relative proportions specified by the probabilities.

The above postulate of the stability of proportionate frequencies in infinite samples cannot be proved, for owing to the limitations of human patience and powers it is impossible to increase the size of an actual sample indefinitely, but on it has been based the whole theory of statistical sampling, the extensive use of which has failed to reveal any inconsistencies that throw doubt on the basic postulate. This will be discussed again in section 2.4.

The infinite population is seen to be distinct from the bulk that is being sampled, and is an abstraction in that its physical existence cannot be shown; it is dependent on the technique of the sampling as well as on the bulk being sampled. Since the interest lies in the characteristics of the bulk, it is important to try and arrange the sampling technique so that the infinite population and the bulk sampled are substantially the same.

2.13. The foregoing discussion of probability may be summed up roughly in the following terms. It is assumed that in the long-run of experience, the proportions of occurrences of different kinds of chance events will tend to have stable values that define the infinite population: The long-run proportion of any one kind of event is its probability. For random sampling, the probabilities of composite events may be calculated from those of simple events by using the laws of mathematical probability. In practice, all probability statements are interpreted in terms of proportionate frequencies in the long-run of experience, even when used to measure a degree of subjective confidence that the event will happen. Thus, an event that happens frequently in the long-run has a high probability, and on any single occasion we have a considerable degree of confidence that an event with a high probability will happen.

We shall now introduce some theoretical frequency distributions that are derived from the laws of mathematical probability, and describe some of the applications to statistical data.

Binomial Distribution

2.2. If we had four perfect dice and threw them together, noting the number of sixes that turned up, we should find at different

throws, four, three, two, one and zero sixes, and if we made enough throws we could form a frequency distribution in which the variate was the number of sixes per throw, or alternatively the number of "not-sixes," and the individuals were throws. This is the distribution dealt with in this section.

The imaginary experiment may be generalised by calling the casting of a six the *occurrence* of an *event* or a *success*, the casting of a number other than a six a *non-occurrence* or *failure*, and the throw of four

TABLE 2.1

Number of Occurrences per Set	Number of Non-occurrences per Set	Proportion of Sets
n	0	p^n
$n-1$	1	$np^{n-1}q$
$n-2$	2	$\frac{n(n-1)}{2!}p^{n-2}q^2$
$n-3$	3	$\frac{n(n-1)(n-2)}{3!}p^{n-3}q^3$
		⋮
0	n	q^n
Total	.. ..	1

dice a *set of n trials* ($n = 4$). Then if the probability of a success is p and that of a failure is q, $(p + q = 1)$, in a set of n independent trials the probability of $(n - s)$ successes and s failures is

$$\frac{n(n-1)\ \ldots\ (n-s+1)}{s!}p^{n-s}q^s* \quad . \quad . \quad . \quad (2.1)$$

* This follows from the rules of mathematical probability given in Section 2.11. The probability that $(n - s)$ particular trials will be successes and s will be failures is $p^{n-s}q^s$ (Rule II). There are

$$\frac{n(n-1)\ \ldots\ (n-s+1)}{s!}$$

ways in which $(n - s)$ successes and s failures may occur in a sample of n (this is a result from the theory of combinations), so from Rule I, the probability that *any* $(n - s)$ trials will be successes and any s will be failures is the expression (2.1).

The probabilities for different values of s are set out in Table 2.1 in the form of a frequency distribution with proportionate frequencies. This is called the *binomial frequency distribution* because the proportionate frequencies are the terms in the expansion of the binomial $(p+q)_{\mathrm{f}}^{n}$. For example, for the four perfect dice the distribution of sixes is the expansion of the binomial $(\frac{1}{6}+\frac{5}{6})^4$ and is shown in Table 2.2. The variate of this distribution is discrete, and it will be noted that all the proportionate frequencies may be described in terms of the two constants p (or q) and n.

TABLE 2.2

Number of Sixes ..	4	3	2	1	0	Total
Proportion of Throws	$\frac{1}{1296}$	$\frac{20}{1296}$	$\frac{150}{1296}$	$\frac{500}{1296}$	$\frac{625}{1296}$	1
Percentage of Throws	0·08	1·54	11·57	38·58	48·23	100·00

It is a matter of algebra to calculate the moments of the general distribution given in Table 2.1; they are:

Mean Number of Occurrences, $l = np$,

Second Moment, $\mu_2 = npq = l\left(1 - \frac{l}{n}\right)$,

whence

$$\textit{Standard Deviation}, \sigma = \sqrt{npq} = \sqrt{l\left(1 - \frac{l}{n}\right)}, \quad \ldots \quad (2.2)$$

Third Moment, $\mu_3 = npq(q-p)$, $\beta_1 = \frac{(q-p)^2}{npq}$,

Fourth Moment, $\mu_4 = npq[1 + 3(n-2)pq]$, $\beta_2 = \frac{1}{npq} + \frac{3(n-2)}{n}$.

It will be noted that the second and fourth moments are symmetrical in p and q, i.e. the result is the same whether the occurrences or failures are the variate, and that for the third moment the interchange of p and q merely alters the sign.

Poisson Series

2.3. A binomial distribution may be imagined in which the probability of a failure, q, is very small, that of a success, p, is nearly equal to unity, and the number of trials per set, n, is exceedingly

large, so that the mean number of failures per set, $m = nq$, is of moderate dimensions. Then for any moderate value, s is negligible compared with n and $(n - s)$ may be equated to n. Doing this in expression (2.1) the probability of s failures is

$$\frac{n^s q^s}{s!}(1 - q)^n = \left(1 - \frac{m}{n}\right)^n \frac{m^s}{s!}$$

and the limit of this as n approaches infinity is

$$e^{-m}\frac{m^s}{s!} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.3)$$

where e is the exponential base $= 2{\cdot}718 \ldots$

The expansion of this for different values of s is the *Poisson Limit to the Binomial* or the *Poisson Series* or the *Law of Small Numbers.*

This distribution is defined entirely by the one constant m, which is the mean. The other moments are given by the limits to equations (2.2) as n approaches infinity, and of these it is only important to note that the second moment, which may be written $(1 - m/n)m$, equals the mean.

To calculate the terms of the series for a given value of m, either the expression (2.3) may be evaluated, with the aid of logarithms, for $s = 0$, $s = 1$, $s = 2$ and so on, or Soper's tables in Pearson's collection of tables (1931) may be used. These tables give the values of expression (2.3) to six decimal places for values of m between $m = 0{\cdot}1$ and $m = 15{\cdot}0$, and within this range, for all values of s that have any proportionate frequency.

Use of the Binomial and Poisson Distributions for Testing Randomness

2.4. One of the most elementary and direct experimental tests of the assumption of an infinite population and of the applicability of the laws of mathematical probability to actual experience consists in seeing if the binomial distribution describes the variations in the numbers of successes in sets of trials. Dice have been thrown by different experimenters many thousands of times in the aggregate, and the resulting frequency distributions have been compared with the corresponding theoretical binomial distributions. Other experiments have taken the form of tossing coins, and thousands of runs of roulette wheels in casinos have been observed for this purpose.

It cannot be said, however, that these experiments have either proved or disproved the laws of probability, for when there has been a discrepancy between theory and experiment, it has usually been possible to find reasons why the particular experiment failed and the general applicability of the laws of probability has remained unquestioned. We believe that these laws are no more to be proved or disproved experimentally than are (say) the terms in a Fourier analysis of a series. They are statements of mathematical relationships that are useful in analysing statistical sampling experience, and do in fact describe an important element in that experience. The justification for the laws is that they are useful in this way. This is the attitude in which the analyses of this section are regarded.

In many collections of statistical data, individuals are of two kinds

TABLE 2.3

Number of Seeds Germinated per Row	Frequency of Rows	
	Actual	Expected
0	6	6·9
1	20	19·1
2	28	24·0
3	12	17·7
4	8	8·6
5	6	2·9
6	—	0·7
7	—	0·1
Total	80	80·0

and divided into sets; the binomial distribution may be used to see if, for any particular collection, the individuals are independent and the two kinds occur at random in the sets. A special experiment to illustrate this was conducted by incubating 800 cabbage seeds on filter paper in rows of ten, and after eight days the number of germinated seeds in each row was counted. These data were formed into a frequency distribution in which the variate was the number of germinated seeds per row and the individuals were rows; this is given in the second column of Table 2.3. If the germinated seeds were distributed at random among the rows, this distribution would

be expected to be a binomial with $n = 10$. We do not know the probability p of a single seed germinating, but may estimate it from the relationship between the mean l, n and p given in equations (2.2). The mean number of germinated seeds per row is $(0 \times 6 + 1 \times 20 + \ldots 5 \times 6) \div 80 = 2{\cdot}175$, and the estimate of p is therefore $2{\cdot}175 \div 10 = 0{\cdot}2175$. The proportionate frequencies of the expected binomial distribution are given by the expansion of $(0{\cdot}2175 + 0{\cdot}7825)^{10}$, and these multiplied by 80 are the expected frequencies of Table 2.3. The agreement between the actual and expected frequencies is quite good,* and as far as may be judged from this limited experience, the variations in germinated seeds from row to row were random; the seeds were well mixed and independent and conditions were uniform, so that there was a constant probability of any one germinating.

It may be objected that since the expected distribution was derived from the actual by choosing p to make the two means equal, the closeness of agreement signifies nothing. This objection is not valid, however, since the two distributions have only been made to agree in two respects, mean and total frequency, and they have not been made to agree in form; that agreement is a consequence of randomness.

Another test of randomness is that the various moments should bear the same relations to each other as those of the theoretical binomial distribution given in expression (2.2). The most important relationship is that between the second moment and mean. The second moment or variance of the actual distribution in Table 2.3, calculated by the technique of paragraphs 1.31 and 1.32, is $1{\cdot}744$, and $l(1 - l/n)$, which may be termed the expected variance, is $1{\cdot}702$; again the agreement is good.

The almost classical example of a Poisson distribution is that given by counts of yeast cells in the squares of a haemacytometer. The liquid in which the yeast cells are suspended can be regarded as consisting of aggregates of molecules of the liquid about equal in size to the yeast cells which are sparsely distributed among them. Then the probability that any aggregate taken at random is a yeast cell (the aggregate being a *trial* and the yeast cell a failure in the language of the theory) is extremely small; but there are very many such aggregates in the liquid under one square in the haema-

* A criterion for judging the closeness of agreement between two distributions will be given in Chapter IV.

cytometer (i.e. in the *set*), so that the mean number of yeast cells per square is finite. Consequently, if the cells are distributed independently and at random through the suspending liquid, the frequency distribution of number per square should be the Poisson Series. Table 2.4 gives the distribution of the counts in 400 cells found by "Student" (1907). The mean number of cells per square is 4·68, and from Soper's tables the Poisson Series having a mean

TABLE 2.4

Number of Cells per Square	Frequency of Squares	
	Observed	Expected
0	—	3·71
1	20	17·37
2	43	40·65
3	53	63·41
4	86	74·19
5	70	69·44
6	54	54·16
7	37	36·21
8	18	21·18
9	10	11·02
10	5	5·16
11	2	2·19
12	2	0·86
13	—	0·31
14	—	0·10
15	—	0·03
16	—	0·01
—	400	400·00

(m) of the same value is constructed, the separate terms being multiplied by 400 to give the expected frequencies of the above table. Thus, for $m = 4·6$, 0·010 052 of the squares should have zero cells and for $m = 4·7$ this frequency is 0·009 095; hence by linear interpolation, for $m = 4·68$ it is $0·010\,052 - 0·8 \times 0·000\,957 = 0·009\,286$, and this multiplied by 400 gives the expected frequency of 3·71. The agreement between the two distributions is quite good. The second moment of the observed distribution is 4·46, and is nearly equal to the mean.

In general, the Poisson distribution would be expected to apply where events occur at random and under constant conditions in a medium (usually space or time) that may be divided into a number of equal zones, provided that relative to the size of the zone the medium is continuous (i.e. there is a very large number of elemental units of medium per zone) and the events are rare "points," i.e. there is room in each zone for an exceedingly large number of events although the actual number in any zone is moderate. Examples are the distribution of microscopic and ultra-microscopic particles and bacteria in liquids, the numbers of α-particles emitted from radio-active substances in intervals of time, and counts of weeds or pests in given areas of land in agricultural field trials.

When the binomial and Poisson distributions fit any given experimental data, it is inferred that the variations are due to chance and cannot be reduced or controlled unless the whole character of the data can be altered by introducing some method of selecting individuals of a given kind. It does not always happen, however, that there is good agreement between theory and experiment. Where there are discrepancies, the plain fact is that the incidence of the event is not random, but the matter is not often allowed to rest there. The investigator usually tries to find the cause of the lack of randomness, and this extends beyond the realm of statistics. Sometimes the cause may be some fault in sampling technique that may be corrected; sometimes the search for the cause may lead to a discovery of scientific value. A frequent kind of discrepancy is that in which the actual variance is greater than the expected, and this is usually interpreted by assuming that conditions are not uniform, regarding the probability of the event as varying from one set of trials or zone to another. Then the actual distribution is composed of several binomial or Poisson distributions superimposed, and the actual variance is equal to the expected variance plus that due to the fluctuations in the probabilities. In such instances, the experience is not merely dismissed as being non-random, but the variation is analysed into two parts, systematic and random. The principles of such analyses will be dealt with in Chapter VI.

There are kinds of discrepancy other than that mentioned, and these lead to other analyses, all of which include a random element.

Readers who are interested in further examples of the uses of the binomial and Poisson distributions are referred to the following

references: Cochran (1936), Fisher (1936*a*), Przyborowski and Wileński (1935), and Tippett (1934).

Normal Distribution

2.5. The Normal distribution may be derived mathematically as another limiting form of the binomial that is approached as n becomes very large, both p and q remaining finite. A binomial distribution may be represented by a histogram with each group centred over the value of the variate corresponding to the number of occurrences per set; there are $(n + 1)$ groups and the outline of the diagram is of the characteristic stepped form. As n, and hence the number of groups, increases, it becomes necessary to reduce the scale of the variate to keep the diagram within reasonable dimensions and so the steps in the outline become smaller. If this process continues indefinitely, it may easily be imagined that in the limit the steps in the outline coalesce to form a smooth curve; this is the frequency curve of the *Normal* or *Gaussian* distribution. Since it is a continuous curve, it necessarily has a continuous variate. Its equation may be written

$$y = \frac{N}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}$$

where e is the exponential base $= 2{\cdot}718\,28\ldots$, x is the variate, and N, m and σ are constants. It will be seen later why the equation is written in this particular form and the $\sqrt{2\pi}$ and $-\frac{1}{2}$ are not incorporated in the constants. The derivation of this equation from the binomial is purely an algebraic process.

This curve is that given in Fig. 2 for $\beta_1 = 0$, $\beta_2 = 3$ and in Fig. 3. It extends between $x = +\infty$ and $x = -\infty$, since only at those extremes does $y = 0$, and it is symmetrical about an ordinate at $x = m$, i.e. about the ordinate at O in Fig. 3.

This frequency curve is deduced from a histogram and consequently areas under the curve and not heights of ordinates represent frequencies. It is therefore appropriate to use the notation and ideas of the integral calculus, to imagine about any given value of x an elemental sub-range dx, and to regard the area under the curve between ordinates drawn at the limits of the sub-range as an element

of frequency, *df* (say). Then, this elemental strip may be regarded as a rectangle of height y and

$$df = ydx = \frac{N}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} dx \quad . \quad . \quad . \quad . \quad (2.4)$$

It is now possible to find the various moments of (2.4). The total area under the curve is the total frequency, and integrating (2.4) between the limits of $x = \pm\infty$ this is found to be N. Hence N in (2.4) is the total frequency. The other moments may be found

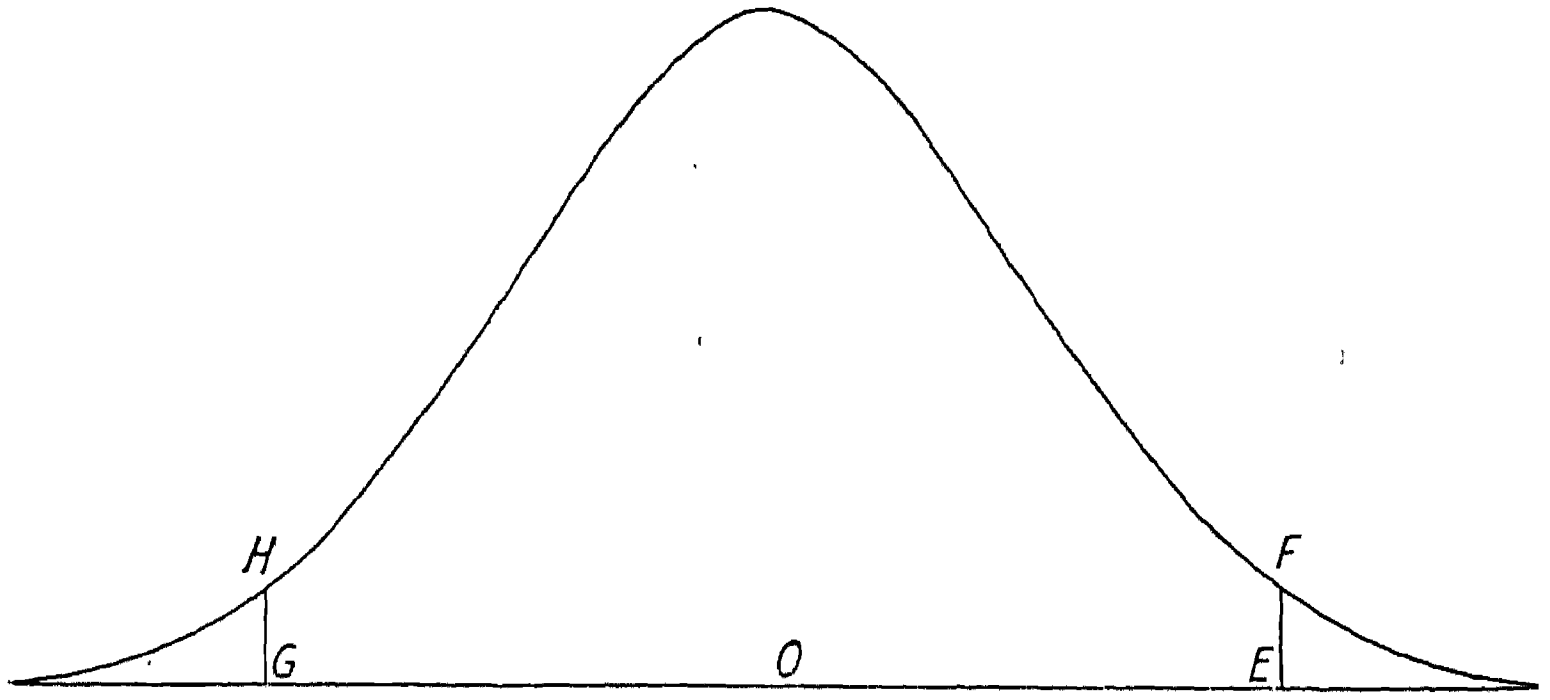


FIG. 3.

from equation (1.3), p. 38, substituting *df* for n_t, the frequency in the sub-group, and integrating instead of summing. Thus the mean is

$$\frac{1}{N}\int_{-\infty}^{+\infty} x\,df = m$$

so that the constant m is the mean of the distribution. Similarly the other moments may be found, and the first four are:

$$\begin{aligned} mean &= m \\ \mu_2 &= \sigma^2 \\ \mu_3 &= 0, \quad \beta_1 = 0 \\ \mu_4 &= 3\mu_2^2, \quad \beta_2 = 3. \end{aligned}$$

These expressions for the mean and variance cannot be deduced from those given in (2.2) for the binomial distribution by writing $n = \infty$, for they would both become infinite; this is balanced in

the derivation of the curve by the ultimate reduction in the scale of the variate, referred to above. The ratios β_1 and β_2, however, are pure numbers, and by putting $n = \infty$ in (2.2) they reduce to the values given above for the normal curve.

Determination of Normal Frequencies

2.51. Apart from the total frequency, N, any given normal distribution is completely characterised by the two constants m and σ. We are interested in calculating proportionate frequencies between various limits of x. This could be done by integrating (2.4) between the limits if it were possible to evaluate the integral, but as it is not, ordinates must be calculated, and the integration performed by quadrature or by some graphical means.

For example, in Fig. 3 the proportionate frequency having values less than a value denoted by G on the variate scale is the area under the curve to the left of GH, expressed as a fraction of the total area under the curve; the proportion between G and E is the area between GH and EF; the proportion greater than E is the area to the right of EF. This process of integration by quadrature is possible, but laborious. However, tables have been calculated, and these reduce the labour considerably.

A complete set of tables for a range of values of m and σ would be enormous, but by performing a mathematical transformation all normal curves reduce to a standard form. The transformation consists in measuring the variate as a deviation from the mean, divided by the standard deviation. If we write

$$w = \frac{x - m}{\sigma}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.5)$$

then for a sample of 1 individual (2.4) becomes

$$df = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2} dw = z dw \quad . \quad . \quad . \quad . \quad (2.6)$$

This may be termed the standardised form of the normal curve, and has a mean of zero and a standard deviation of unity. The *probability integral* at w is the area under the curve to the left of an ordinate drawn at w and may be written

$$A_w = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2} dw.$$

Very complete tables of A_w and z for positive equally spaced values of w have been calculated by Sheppard and are included in *Tables for Statisticians and Biometricians*, Vol. I (Pearson, 1931), where our w is called x and our A_w is called $\frac{1}{2}(1 + \alpha)$.* To find the probability integral for a negative value of w, use is made of the fact that the distribution (2.6) is symmetrical about an ordinate at $w = 0$. If A_{-w} is the integral at $-w$ and A_w is the integral at w, it is easy to see that

$$A_{-w} = 1 - A_w \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.7)$$

There are also tables that give values of w for equally spaced values of A_w while others give w for equally spaced values of $2(1 - A_w)$. If OE in Fig. 3 represents a deviation $+ w$ and OG a deviation $- w$ ($OE = OG$ and since O is at the centre of symmetry it represents $w = 0$) the area to the left of EF is A_w and the "tail" to the right is $(1 - A_w)$. Similarly, by symmetry the area of the "tail" to the left of HG is $(1 - A_w)$ and so $2(1 - A_w)$ is the sum of the two "tails" beyond the deviations $+ w$ and $- w$. The probability integral is sometimes tabulated in this form because it is of interest in sampling theory; this is the method of presentation used by Fisher (1936*a*).

To find the probability integral for any value of an actual variate (x in our notation) it is transformed to w by equation (2.5) and the corresponding integral is that required. Thus, if $m = 67{\cdot}697\,6$ inches and $\sigma = 2{\cdot}704\,8$ inches and it is required to find the probability integral at $x = 60{\cdot}5$ inches say, then

$$w = \frac{60{\cdot}5 - 67{\cdot}697\,6}{2{\cdot}704\,8} = -\,2{\cdot}661\,0$$

From Sheppard's tables the value of the integral at a deviation of $2{\cdot}66$ is $0{\cdot}996\,09$ and at $2{\cdot}67$ it is $0{\cdot}996\,21$; so the first difference is $0{\cdot}000\,12$, and by linear interpolation the integral for $w = +\,2{\cdot}661\,0$ is

$$A_w = 0{\cdot}996\,09 + 0{\cdot}000\,12 \times 0{\cdot}10 = 0{\cdot}996\,10.$$

Hence, from equation (2.7), the integral for $w = -\,2{\cdot}661\,0$ is $0{\cdot}003\,90$.

The proportionate frequency between any two values of the variate may be found by taking the difference of the two corre-

* In this book we shall continue to use our own notation, even when referring to these tables.

sponding integrals. Thus, in the above example the integral corresponding to $x = 59{\cdot}5$ is found to be 0·001 22 and the proportionate frequency between 59·5 and 60·5 inches is 0·003 90 − 0·001 22 = 0·002 68.

A normal frequency distribution may be "fitted" to an actual distribution by putting m and σ equal to the computed mean and standard deviation respectively and then finding the frequencies in the sub-groups from the normal probability integrals. The proportionate frequency calculated in the above example is for the second group of the distribution of heights of fathers in Table 1.5 and has been calculated in this way; the process is completed in the later columns of that table. Column (8) gives values of w corresponding to the limits of the sub-ranges, column (9) gives probability integrals, in column (10) these are converted to frequencies by multiplying by the total $N = 1\,078$, and the normal or "expected" frequencies in column (11) are the differences of the values in the previous column. These may be compared with the actual frequencies, n_t, in column (3). In order to plot the curve we must find the ordinates. Sheppard's tables give values of z, the ordinates of the standardised curve corresponding to the deviations w (see equation 2.6), and these are given in column (12) of Table 1.5. The ordinates of the actual curve are obtained by the transformation

$$y = \frac{zN}{\sigma} = 398{\cdot}55\, z,$$

and these are in column (13). In Fig. 4 the curve drawn from these ordinates is superimposed on the histogram.

Sometimes a frequency or proportionate frequency is given, and it is desired to find the corresponding value of the variate. The value of the standardised variate w may be found from Sheppard's tables, and hence, knowing m and σ, the actual variate be calculated from equation (2.5).

For example, suppose it is required to know for the data of Table 1.5 the limit of height such that 20 per cent of the fathers are shorter than the limit and 80 per cent are higher, assuming the distribution to be normal with mean and standard deviation equal to the values already computed. Then $A = 0{\cdot}2$, and since this is less than 0·5 it corresponds to a negative value of w; we must

therefore first find w corresponding to $A = 1 - 0{\cdot}2 = 0{\cdot}8$. From Sheppard's tables, the value of the variate at which $A = 0{\cdot}799\,546$ is $0{\cdot}84$ and the value at which $A = 0{\cdot}802\,338$ is $0{\cdot}85$. Hence, by linear interpolation, the value at which $A = 0{\cdot}8$ is

$$w = \frac{0{\cdot}8 - 0{\cdot}799\,546}{0{\cdot}802\,338 - 0{\cdot}799\,546} \times 0{\cdot}01 + 0{\cdot}84 = 0{\cdot}841\,63.$$

Hence the value of w at which $A = 0{\cdot}2$ is $-0{\cdot}841\,63$ and if this is substituted in (2.5) the limit of height is found to be $65{\cdot}421$ inches.

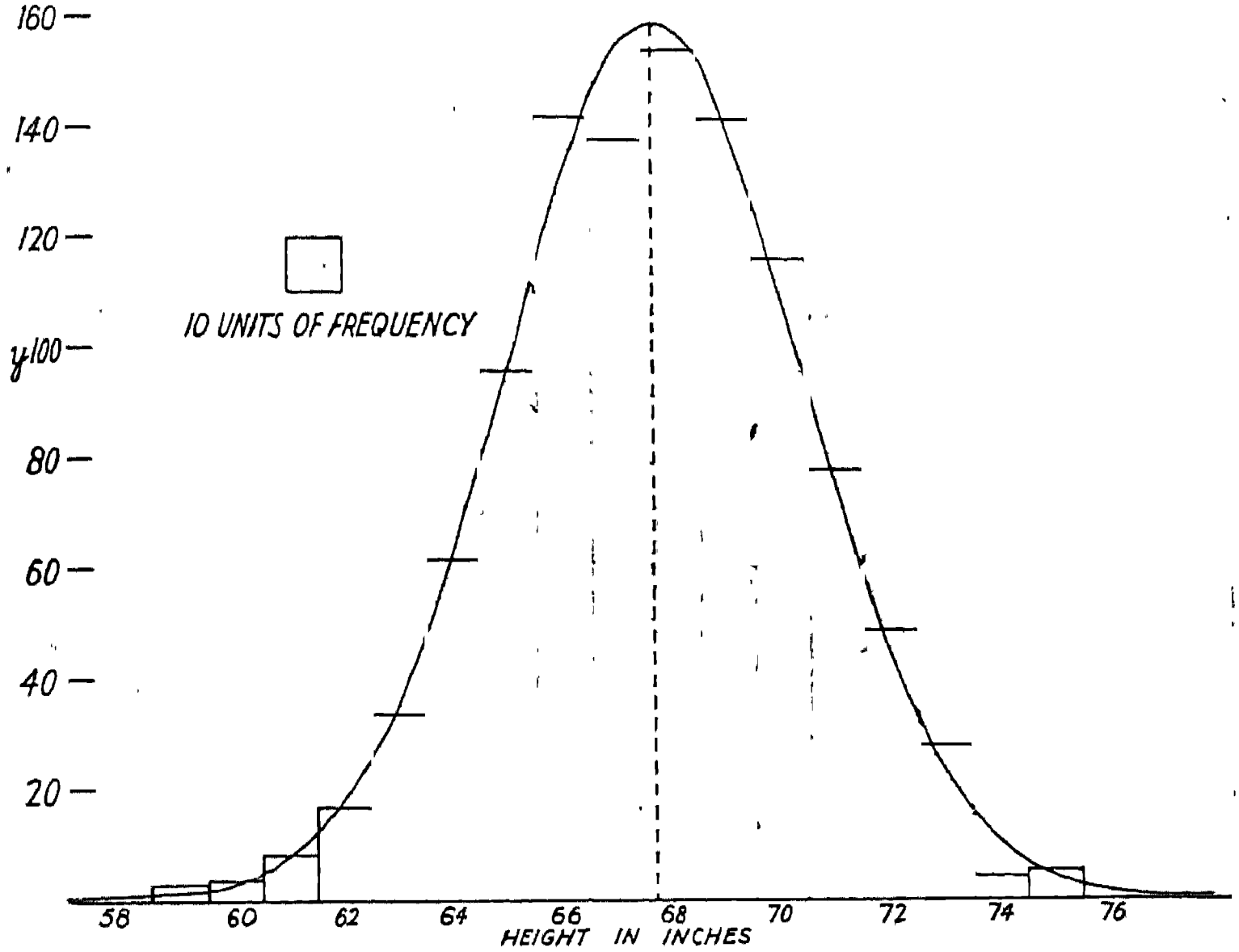


FIG. 4.

Relation between Normal Frequencies and Standard Deviation

2.52. We cannot here give a full table of the probability integral of the normal distribution, but a few important values are given in Table 2.5. The first entry in this table is obvious; the ordinate at $w = 0$ is the axis of symmetry of the curve, and the area up to this ordinate is half the total area.

The second entry of the variate is the quartile deviation; the

proportionate frequency between positive and negative values of this deviation is half the total. The ordinates *GH* and *EF* drawn in Fig. 3 are at values of the variate corresponding to $w = +2{\cdot}0$ and $-2{\cdot}0$, and the proportionate area of the two "tails" beyond these limits is 0·045 50; i.e. about 5 per cent of the individuals in a normal population deviate from the mean by twice the standard deviation or more. Similarly, from the last entry in Table 2.5 we see that 99·73 per cent of the individuals are contained within a total range of six times the standard deviation. These data will assist readers in appreciating the significance of the standard deviation as a measure of variability when applied to a distribution that is approximately normal.

The proportional relations between the standard deviation and

TABLE 2.5

NORMAL DISTRIBUTION

Variate w	Probability Integral A_w	Sum of Two "Tails" $2(1-A_w)$
0	0·500 00	1·000 00
0·674 49	0·750 00	0·500 00
1·0	0·841 34	0·317 31
2·0	0·977 25	0·045 50
2·6	0·995 34	0·009 32
3·0	0·998 65	0·002 70

other measures of dispersion given above in section 1.23 are true only for normal distributions, although they are roughly applicable also to many distributions that are approximately normal.

Practical Applicability of the Normal Distribution

2.53. The Normal distribution is a continuous curve, and first we must discuss the applicability of frequency curves in general to practical data.

It is assumed that for most infinite populations with a continuous variate, the frequency distribution may be represented by a continuous frequency curve. This is an extrapolation of the practical experience that as the size of the sample is increased, the sub-ranges of the distribution may be reduced and the outline of the histogram

usually becomes more regular. Where a form of curve can be assumed, the parameters may be estimated from a large sample, and a curve be fitted to the actual distribution, as has been done for Table 1.5.

The normal curve has also been deduced mathematically as the distribution that results from the combination of an infinite number of very small random errors, and this fact together with the fact that the distribution is a special case of the binomial gives it a fundamental status. It is believed by some to represent the deviations due to experimental errors in measurements of a physical constant, and a quantity that is distributed according to this law is sometimes held to be subject only to chance causes that cannot be controlled. Conversely, where a distribution is skew, it is often inferred that superimposed on the chance variations are some larger ones due to a few important causes that may be controlled. We are dubious of this use of the normal distribution. Doubtless normality may be made a definition, and hence a test, of "randomness" just in the way that conformity to the binomial and Poisson distributions has been interpreted, but it is doubtful if for the normal curve such a course has any practical significance. Frequently, when man has done all he can to control the variation in a character and the remaining variations are practically random, the resulting distribution is far from normal. Experimental errors in physical measurements are by no means always normally distributed, and some quantities like the strength of elements of various natural and manufactured materials have essentially skew frequency distributions. It may also happen that the distribution of a quantity is to all intents and purposes normal, and yet a further degree of control may be possible. On the whole, we doubt if the normal distribution is of more than empirical value; it is a convenient means of describing any data it happens to fit.

In presenting this view of the place occupied by the normal curve in the statistical scheme, it would be a mistake to underestimate its importance. The curve does in fact represent closely a very large number of experimental distributions and approximately represents many more. It is important, too, because it is the distribution on which most of the statistical theory of errors is based.

Non-normal Curves

2.6. There are several systems of smooth curves for describing data which do not follow the normal law, of which Pearson's is

probably the most useful; all the curves in Fig. 2 are of this system. The type is decided upon from the values of the two β ratios, and then these, together with the standard deviation and mean, are used to find the unknown parameters in the equation. It is not often in practical work, however, that any useful purpose is served by fitting such curves (their chief use is in sampling theory and in actuarial work), so we will merely refer the interested reader to some such book as Elderton's *Frequency Curves and Correlation*.

Sampling Distributions

2.7. A sampling distribution results if a large number of finite random samples are taken from an infinite population, some single

TABLE 2.6

Frequency Distributions

	27–	30–	33–	36–	39–	42–	45–	48–	51–
Individual observations ..	4	2	1	3	7	7	10	17	13
Means of 5	—	—	—	—	—	3	4	5	2

	54–	57–	60–	63–	66–	69–	72–	Total
	13	8	7	3	1	3	1	100
	5	1	—	—	—	—	—	20

characteristic of each sample is computed and the computed values are formed into a frequency distribution. For example, the observations of Table 1.3 may be regarded as forming twenty samples of five observations from one population. The means of these samples are given in Table 1.3, and are formed into a frequency distribution in Table 2.6. This table also gives the distribution of the individual observations.

The distribution of means is called *sampling distribution* of the mean. Like any other frequency distribution it has a mean, and a standard deviation. The standard deviation of such a distribution is called the *standard error* of the mean.

Other constants of the samples of five could have been calculated,

for example the standard deviations or the medians. Each of these would have had its own sampling distribution and standard error.

Most sampling distributions in common use may be deduced mathematically by applying the laws of probability, but given sufficient time and energy one could determine them experimentally by making up an artificial population (say) by writing numbers on cards, shaking up in a bag, drawing samples and calculating the means or other statistical constants much in the way in which Table 2.6 was constructed from Table 1.3, except that the scale of the experiment would need to be much larger. This kind of technique is actually used when the distribution cannot be deduced mathematically.

Sampling Distribution of Mean

2.71. We are particularly concerned here with the sampling distribution of the mean in samples of size N drawn from an infinite population in which the individuals are normally distributed. This is itself a normal distribution with a mean equal to the mean of the individuals in the population and a standard error of

$$\frac{\sigma}{\sqrt{N}}$$

where σ is the standard deviation of the individuals.

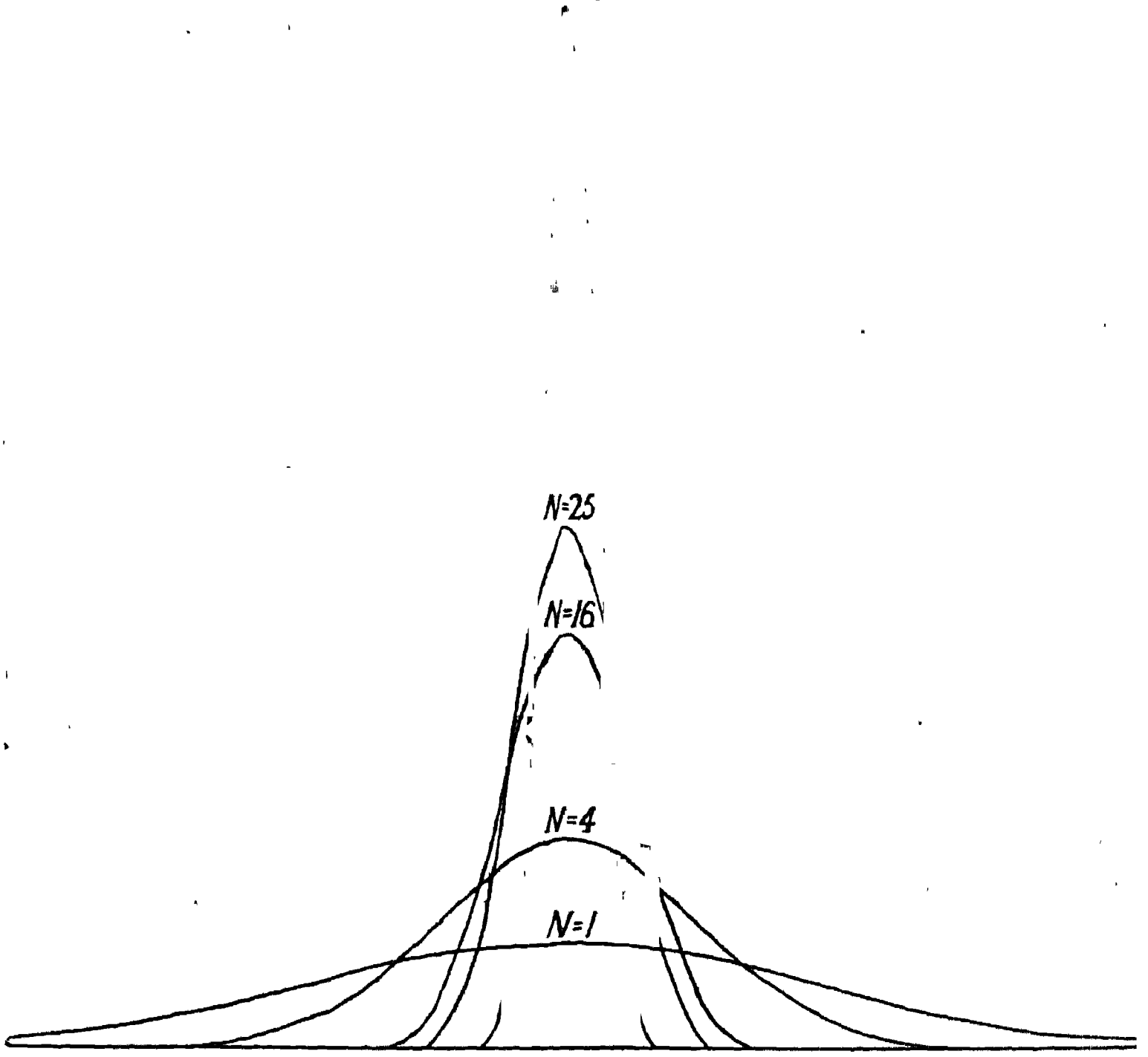
It will be noticed that as N increases, the standard error of the mean decreases. This is shown in Fig. 5, where there are given the sampling distributions of the means of samples of 1, 4, 16, 25 and 100 from a population in which the individuals are normally distributed with an unspecified standard deviation. This population distribution is that in Fig. 5 for $N = 1$. For the larger samples, there is less dispersion about the population mean. This is the statistical demonstration of the common experience that a large sample gives a more accurate representation of the population than a small one; the increase in precision is measured by a reduction in standard error.

The normal probability integral may be used to calculate proportionate frequencies of samples having means within given limits, and Table 2.5 may be used by reading "deviation of sample mean from population mean, divided by standard error" for "variate w." Thus only 0·045 50 or about 1 in 20 of the possible samples have

means that differ from the population value by more than twice the standard error.

Deduction of Sampling Distributions of Mean and Standard Deviation

2.72. The sampling distributions of both the mean and standard deviation may be deduced together. The proof is taken from a paper by Irwin (1931).



NORMAL SAMPLING DISTRIBUTIONS

FIG. 5.

Let the population mean be $\bar{\xi}$ and the other particulars as stated above. Also let the variate be x and the values in any one sample of N be $x_1\, x_2 \ldots x_s \ldots x_N$. Since the population distribution is continuous, we cannot state what is the probability of a value x_s, for an ordinate of the frequency curve drawn at x_s has no area, but

the probability of a value lying within an elemental range dx_s about a value x_s is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_s-\bar{\xi})^2}{\sigma^2}} dx_s.$$

Applying the second rule of probability, the probability of a sample having values lying within ranges $dx_1\, dx_2 \ldots$ of $x_1\, x_2 \ldots$ which we may shortly describe as "the probability of the sample," is

$$\frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} e^{-\frac{1}{2}\frac{(x_1-\bar{\xi})^2+(x_2-\bar{\xi})^2+\ldots(x_N-\bar{\xi})^2}{\sigma^2}} dx_1 dx_2 \ldots dx_N, \quad \text{or}$$

$$\frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} e^{-\frac{1}{2}\frac{x_1^2+x_2^2+\ldots x_N^2-2\bar{\xi}(x_1+x_2+\ldots x_N)+N\bar{\xi}^2}{\sigma^2}} dx_1 dx_2 \ldots dx_N.$$

Now if $\bar{x}$ and s are the mean and standard deviation in the sample,* according to sections 1.23 and 1.31,

$$x_1 + x_2 + \ldots x_N = N\bar{x} \quad \text{and} \quad x_1^2 + x_2^2 + \ldots x_N^2 - N\bar{x}^2 = Ns^2$$

Substituting these in the exponent of the above expression, the probability of the sample is

$$\frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} e^{-\frac{N}{2}\cdot\frac{s^2+(\bar{x}-\bar{\xi})^2}{\sigma^2}} dx_1 dx_2 \ldots dx_N.$$

Many of the possible samples from this population will have a mean value $\bar{x}$ and a standard deviation s, and to find the probability of a sample with these two constants lying within ranges $d\bar{x}$ and ds of $\bar{x}$ and s we must apply the first probability rule and integrate the above expression for all samples having these two constants. This is a mathematical step involving a transformation to polar co-ordinates in N-dimensional space and a subsequent integration, and leads to the following expression for the probability of a mean of $\bar{x}$ and a standard deviation of s:

$$df = Ks^{N-2} e^{-\frac{N(\bar{x}-\bar{\xi})^2}{2\sigma^2}} e^{-\frac{Ns^2}{2\sigma^2}} d\bar{x}\, ds$$

* For the first time it is necessary to distinguish clearly between the population value of a statistical constant and the value estimated from a sample. Usually we shall denote population values by Greek letters and the sample values by corresponding italic letters. This is why we have used $\bar{\xi}$ instead of m in the equation for the normal distribution.

where K is a constant. This is the probability or proportionate frequency distribution of the two statistical constants. Since the term containing $\bar{x}$ does not contain s, and *vice-versa*, the converse of the second rule of probability implies that $\bar{x}$ and s have independent probabilities, and the two distributions may be written separately:

$$\left.\begin{aligned} df &= K_1 e^{-\frac{N}{2}\frac{(\bar{x}-\bar{\xi})^2}{\sigma^2}} d\bar{x} \\ \text{and} \qquad df &= K_2 s^{N-2} e^{-\frac{Ns^2}{2\sigma^2}} ds. \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.8)$$

By integrating the first expression over the whole range and equating the result to unity, the constant K_1 is found to be

$$K_1 = \frac{\sqrt{N}}{\sqrt{2\pi}\,\sigma}.$$

The distribution of the mean is a normal one with mean equal to $\bar{\xi}$ and standard deviation equal to $\sigma/\sqrt{N}$. That of the standard deviation is more complicated.

CHAPTER III

ERRORS OF RANDOM SAMPLING AND STATISTICAL INFERENCE

MUCH statistical work is concerned with attempts to learn something of the characteristics of populations from measurements made on finite representative samples. It is a fact of experience that successive samples from one population usually differ amongst themselves, and therefore in general they differ from the population. They are subject to the so-called errors of sampling that bring uncertainty into any inferences that may be made regarding the population. This chapter is concerned with the theory of statistical inference, taking account of these errors; the applications of the theory are described for large samples only.

RANDOM AND REPRESENTATIVE SAMPLES

3.1. We shall deal only with simple random samples, in which, as stated in section 2.12, every individual is independent of every other. A sample of 1 000 men consisting of 500 pairs of brothers is not random, for there is a tendency for brothers to be alike and there are only 500 independent individuals; the others are related, statistically as well as by blood, to the first 500. Special precautions are necessary in practice to obtain satisfactory samples; bales of cotton, sacks of corn and the like are seldom uniform, and a small quantity taken from one place is not representative of the whole. In such circumstances, it is necessary to select the individuals one at a time, but where the bulk is well mixed and statistically homogeneous, a group of individuals taken from any part is effectively a random sample.

The theory of sampling describes the relations between samples and the infinite population, and this last will differ from the actual population or bulk if the sample is biased. In practice, it is the bulk that is of interest, and it is desirable to make the sample a representative one by seeing that every individual in the bulk has an equal chance of being included. This is not always easy. For example, when taking cotton hairs singly from a bale, there is an unavoidable tendency to take too many of the longer hairs. The method of drawing individuals must be independent of their character.

The degree of elaboration necessary in a sampling technique to ensure randomness and representativeness depends on the nature of the material being sampled, and the arrangement of a technique is more a matter for the experimentalist than for the statistician. Readers are advised, however, that investigation in particular fields has often shown that much more care in sampling is required than appeared to be necessary at first sight. Although it is impossible to give general rules, the following description of two particular sampling methods may be useful as illustrations.

The first method is that applied to sampling cotton hairs. Hairs are taken from different parts of the bale, not singly but in tufts sufficiently large to avoid the bias towards length. Each tuft is then divided roughly into two halves and one half is discarded. The other is halved again, one half is discarded and the other is again divided. This is repeated until the final reduced tuft contains only a few hairs. A number of such reduced tufts is combined to form the final sample, every hair of which is measured. If at each division the half to be discarded is decided by lot, say by the toss of a coin, the final sample will be truly representative. But it will not be a simple random sample unless the final reduced tufts contain only one hair each. However, we shall show how to treat such complex samples in section 10.11.

The second method may often be applied when the population to be sampled is large but finite. Suppose we wish to inspect a random sample of houses in a large town. The town may be divided into wards, and the wards into streets. One ward may be chosen at random, a street may be chosen at random from those in the ward, and a house from those in a street; this house is a randomly chosen individual, and if the whole process is repeated several times, the resulting houses are a random sample of the town, provided the wards and streets are approximately equal in size. If the wards, the streets within each ward, and the houses in each street are all systematically numbered, the problem of random selection resolves itself into one of choosing a number at random. This process is facilitated by Random Sampling Numbers (Tippett, 1927), which is a collection of 40,000 numbers from 0 to 9, arranged in random order. By combining an appropriate number of digits taken anywhere from this collection, random numbers of almost any magnitude may readily be obtained.

It should be noted that if arrangements are made to represent

each ward equally in the sample or to see that no two houses are in the same street, the sample is representative but not random. Such may be called a stratified sample, and its random errors are dealt with in section 10.13.

The theory of random sampling is sometimes applied to experimental errors in physical determinations. It is not unreasonable to regard the limited number of measurements that may actually be made of a physical quantity as a random sample of an infinite population of measurements that could conceivably be made under the same conditions. The average of this population, however, is not necessarily the true value of the quantity; it is affected by errors of bias due to the particular conditions of the experiment, to such factors as the personal error of the observer and idiosyncrasies of the apparatus. Since in most physical experiments errors of bias are liable to be of the same order of magnitude as random errors, the theory of random sampling is of very limited utility in this field.

The qualities of randomness and representativeness do not depend on the size of the sample; a sample of five, if properly chosen, is as random and representative as one of five thousand. In this chapter we are limiting, not the general theory, but the application of the theory to large samples, i.e. samples of at least a few hundreds, and this is done to simplify matters. The more exact theory for small samples will be given in Chapter V.

We shall again make the further assumption that the population sampled is infinite in the sense that its composition is not appreciably altered by the abstraction of the sample. This is satisfied, practically, if the individuals are drawn one at a time from even a very small bulk provided they are returned to the bulk and well mixed with the other individuals between each draw.

Tests of Significance

3.2. Many scientific investigations involve the employment of the method of framing working hypotheses and testing them experimentally. As long as the experiments fail to disprove them, so long are the hypotheses accepted. This is the general method by which many statistical inferences are made. A hypothetical population of certain characteristics is postulated, and if the sample is such that it could reasonably have come from that population, the hypothesis is accepted. Owing to sampling errors, however, there is no sharp dividing line between samples that could have come from the

hypothetical population and those that could not. It is only possible to give a probability that a sample like the one observed could have come from the population. If the probability is low, the hypothesis is rejected; if it is high, the hypothesis is accepted and the deviation between the sample and postulated population is attributed to errors of sampling. The sample may be characterised by one or more statistical constants, and in the next paragraph the process will be illustrated for the mean of samples.

If there is some ground for assuming* the population to be normal with a given mean and standard deviation, the distribution of the means of samples of any given size is the sampling distribution described in section 2.71, and from the probability integral, the probability of a sample mean deviating from the population mean by more than any value may be deduced. For example, we see from Table 2.5 that the probability of a sample mean differing from the population mean by plus or minus three or more times the standard error is 0·002 70. This is low, and if any actual sample does differ in mean from the population value by more than this amount, we infer that the deviation may not reasonably be attributed to random errors and the hypothesis regarding the population has to be abandoned.

The probability of a random deviation exceeding any given value is called the *level of significance* of the deviation, and is often expressed as a percentage. For example, a deviation of ± 3 times the standard error is on the 0·27 per cent. level of significance and one of 1·0 times the standard error is on the 32 per cent. level.

By way of example we will assume the lengths of 4 000 hairs of an Indian cotton given by Koshal and Turner (1930) to be an infinite population.† Their mean length is 2·33 cm. and the standard deviation is 0·480 6 cm. The first thousand hairs were selected by a different method from the rest and gave a mean of 2·54 cm. Is this deviation compatible with the hypothesis that the 1 000 are a random sample from the 4 000 and that the difference in means is due to random errors, or is the difference large enough to indicate

* There is no fundamental distinction between an hypothesis and the assumptions (e.g. normality). However, a good test is sensitive to falseness in the hypothesis, which is what we want to test, and comparatively insensitive to errors in the assumptions, which are subsidiary.

† This assumption is an approximation, for 4 000 is not infinitely large compared with 1 000, the size of the sample we are testing.

that the change in technique has had an effect? The standard error of the mean is

$$\frac{0{\cdot}480\,6}{\sqrt{1\,000}} = 0{\cdot}015\,2,$$

and the deviation of 0·21 cm. is over 13 times the standard error. Sheppard's normal probability tables show that this is far beyond the 0·000 001 level of significance, and the hypothesis is untenable.

Any deviation that is large enough and is on a sufficiently low probability level to lead to a rejection of the hypothesis regarding the population is said to be *statistically significant* and the mean of the sample is said to be significantly different from that of the population. The question arises: at what probability level does a deviation become significant? There is no rational probability level at which possibility ceases and impossibility begins, but it is conventional to regard a probability of 0·05 as the critical level of significance. The considerations that govern the choice of this level will be discussed in section 3.3, and it is sufficient here to state that this convention has been found to give a satisfactory rule of action in most circumstances. It will be seen from Table 2.5 that a deviation of twice the standard error corresponds roughly to a probability level of 0·05, so we have the following working rule:

> A deviation of a sample mean from an assumed population value of twice the standard error lies on the 0·05 (or 5 per cent.) level of significance, and a deviation greater than this amount is statistically significant.

As a measure of dispersion of the sampling distribution of sample means, the quartile deviation is sometimes used instead of the standard error, and is called the *probable error*. The probable error is 0·674 49 times the standard error, and three times the probable error is roughly equivalent to twice the standard error. The probable error has no particular advantages, and as it involves the troublesome factor 0·674 49, it is rapidly going out of use.

Significance of Difference between Two Sample Means

3.21. The most common situation is one in which the investigator has two samples, and wishes to know if their differences are real or may be attributed to errors of random sampling. Again we shall confine attention to the two means. The appropriate hypothesis is

that the two samples are from populations having the same mean, and the probability level of the observed difference is calculated accordingly. Again the hypothesis is accepted if the level is fairly high and there is no statistically significant difference between the means; if the level is low (say below 0·05), the hypothesis is rejected and the difference is real. To calculate these probabilities it is necessary to know the sampling distribution of the difference between two means.

Let the total numbers of individuals in the two samples be N_1 and N_2. Then it is possible to imagine a sampling experiment in which a very large number of pairs of samples are taken from the respective populations, the number of individuals in the first of each pair being N_1 and the number in the second being N_2. For each sample the mean may be found, and hence for each pair, the difference between the two means. There will be as many differences as there are pairs of samples, and these may be formed into a frequency distribution—the sampling distribution of the difference between two means. This distribution has been deduced mathematically and is normal with a mean value of zero, as may be expected, and a standard deviation (or standard error) larger than the standard error of either sample mean taken separately. If SE_1 is the standard error of the distribution of one series of means and SE_2 is that of the other, the standard error of the distribution of differences is

$$SE_{1-2} = \sqrt{(SE_1)^2 + (SE_2)^2}* \quad . \quad . \quad . \quad (3.1)$$

provided the two samples are independent. Further, if σ_1 and σ_2 are the standard deviations of the individuals in the two populations, the standard error of the difference between the two means is

$$SE_{1-2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \quad . \quad . \quad . \quad . \quad . \quad (3.2)$$

Usually, σ_1 and σ_2 do not differ appreciably, and it is reasonable to include in the hypothesis the postulate that the two populations are the same, so that $\sigma_1 = \sigma_2$.

In these circumstances, if the two single samples are of the same size they have the same standard error, and the "twice the standard error" criterion leads to the working rule that differences greater

* This follows easily from the equations in section 3.55.

than three times the standard error of a single mean are significant; for if $SE_1 = SE_2$, $SE_{1-2} = \sqrt{2}SE_1$ and $2\sqrt{2}$ is nearly equal to 3.

In order to compute the standard error, it is necessary to know σ, the standard deviation of the individuals in the population. Frequently this is unknown, and as an approximation an estimate s obtained from the samples is used instead. This practice limits the application of the theory to large samples. Where there are two samples and a common standard deviation is assumed, some combined estimate of s should be used since the hypothesis is that the samples are from the same population. If s_1 and s_2 are the two sample estimates, a suitable combined estimate is

$$s = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}} \quad . \quad . \quad . \quad . \quad . \quad (3.3)$$

and substituting this value of s for the values of σ in equation (3.2) we have

$$SE_{1-2} = \sqrt{\frac{s_2^2}{N_1} + \frac{s_1^2}{N_2}} \quad . \quad . \quad . \quad . \quad . \quad (3.4)$$

Frequently, however, the standard errors of separate means are calculated more or less as a routine and it is convenient to use these directly in equation (3.2) leading to the expression

$$SE_{1-2} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \quad . \quad . \quad . \quad . \quad . \quad (3.5)$$

This latter course is not consistent with the hypothesis that the samples are from the same population, but when the samples are large the two courses usually lead to practically the same results.* The second is used in the following example.

The British Association Report for 1883 gives on p. 256 distributions of the heights of men born in England and Scotland; as an example we will test the significance of the difference between the two means. The necessary data are in Table 3.1.

The difference in height is 1·108 1 inches, and its standard error is

$$\pm\sqrt{0{\cdot}032\,38^2 + 0{\cdot}068\,68^2} = \pm\, 0{\cdot}075\,9;$$

* The second course is consistent with the original hypothesis that the samples are from two populations with the same mean and different standard deviations. This is a legitimate assumption, but is not usually as acceptable on other grounds as the one given above.

1·108 1 is 14·6 times 0·075 9, and we thus conclude that Scotsmen are really taller than Englishmen.

General Discussion

3.3. We shall first discuss the choice of the level of statistical significance. If we work consistently according to the rule that the hypothesis is acceptable provided the observed difference or deviation is below a given level and is rejected if the deviation is above, one of four possible situations may arise*:

We may be in error because we:

(1) Reject a true hypothesis, or
(2) Accept a false one;

TABLE 3.1

Country	Number in Sample	Mean Height (inches)	Standard Deviation (inches)	Standard Error of Mean
England ..	6 194	67·437 5	2·548	$2{\cdot}548 \div \sqrt{6\ 194} = \pm 0{\cdot}032\ 38$
Scotland ..	1 304	68·545 6	2·480	$2{\cdot}480 \div \sqrt{1\ 304} = \pm 0{\cdot}068\ 68$

or we may be correct because we:

(3) Accept a true hypothesis, or
(4) Reject a false one.

In accepting a hypothesis according to the rule, we only do so tentatively, since no hypothesis is, or ever can be, finally proved.

In the long run of statistical experience, the ratio of wrong inferences under (1) to total inferences under (1) and (3) when the hypothesis is a true one can be made as low as we please by making the level of significance high†; indeed, this ratio is the probability level, and the rule already proposed leads to a true hypothesis being rejected once in every twenty experiments for which a true hypo-

* This analysis was given by Neyman and Pearson (1928), and some writers have accepted the classification to the extent of referring to situations (1) and (2) as "errors of the first and second kinds."

† A low probability corresponds to a high level of significance.

thesis is made. By adopting a level of 0·01, this kind of error is made only once in every hundred experiments. Unfortunately, by using a higher level of significance, we are more likely to accept the hypothesis and hence are more likely to be wrong under situation (2). Thus the choice of the level for a criterion of acceptance or rejection is a compromise between the risks of the two types of error.

Unfortunately, it is not possible to state the proportion of erroneous inferences under (2) to the total experiments in which a false hypothesis has been made. It will depend to a considerable extent on how near the hypothesis is to the truth as well as on the stringency of the test.

The choice of the critical level of significance will depend to some extent on circumstances, but it is governed largely by the following consideration. We usually have a large body of knowledge and a scientific tradition to guide us, so that there is a fairly strong predisposition towards accepting any hypothesis we make. One of the strongest traditions favours simple hypotheses involving few constants in preference to complex ones involving many constants. For example, in investigating a possible difference in height between Englishmen and Scotsmen as in paragraph 3.21, it was in accordance with scientific practice to prefer the hypothesis of a single mean height to one involving two heights, i.e. to assume no difference until the contrary was proved. It is because of this, and because the effect of an error under (2) is less serious than one under (1) that the chosen criterion of significance usually has a probability level as low as 0·05, and in instances where there are exceptionally strong *a priori* reasons for preferring the hypothesis or an error under (1) may have important consequences, the appropriate probability level is 0·01 or even lower. On the other hand, most hypotheses tend to be of a negative character in that they are in accordance with existing knowledge, and by choosing too low a probability corresponding to a level of significance that is too high, the proportion of mistaken inferences of the second kind may be too great, and advance of knowledge may be unjustifiably impeded. Too much scepticism may be obstructive.

3.31. It can be seen that one means of reducing the effect of erroneous inferences of the second kind without affecting those of the first is by reducing the standard error of the statistical constant under test. For a given level of significance, a smaller standard error

leads to a correspondingly smaller deviation lying just on the chosen level of significance, and so to a smaller deviation wrongfully discounted as not significant. The standard error may be reduced by increasing the size of the sample, and we shall see in section 3.7 that from this point of view some statistical constants are preferable to others that measure equivalent properties of the frequency distribution.

The statistical significance gives no information as to the magnitude or practical importance of any difference; that can only be judged by one with technical knowledge of the subject to which these methods are applied. A very large sample may make very small and unimportant differences overwhelmingly significant, while if the sample is small, large and important differences may be obscured by random errors. The verdict "not significant," therefore, is more like the "not proven" of Scots law than "not guilty." If an approximate value of the standard deviations of the populations is available (say from preliminary samples), it is always possible to estimate how large samples are necessary to show up a given difference; N should be amply big enough to make twice the standard error of the difference less than the specified one.

If a sample mean is significantly different from a hypothetical population value or the difference between two sample means is significant, statistical theory can shed no light on the cause of the deviation. It may be either that the bulk sampled is really different from the hypothetical population or that the sample is not truly representative and random; either that there is a real difference between the two populations sampled or that there is some difference in sampling technique.

It may be noted that there is nothing in these tests of significance that provides a criterion for choosing between hypotheses that may be compatible with the data. Such a choice must be made on non-statistical grounds.

3.32. So far we have taken as the 0·05 level of significance a positive or negative deviation for which the two "tails" of the probability curve total 0·05, i.e. each "tail" separately is 0·025. This is because we usually have no *a priori* reason for deciding that the deviation or difference should have one sign or another. In the example of

section 3.2 treating of the lengths of cotton hairs, the mean of the sample of 1 000 happened to be greater than the assumed population; we should have tested the deviation in the same way had the mean been less. In the example of section 3.21 the Scotsmen were apparently taller, on the average, than the Englishmen; we should have tested the difference had the Scotsmen been shorter. In circumstances like these, the probabilities must be added for both positive and negative deviations or differences, so as to make the calculation correspond to practice and to give correctly for the long-run of experience the ratio of mistaken to total inferences under situations (1) and (3) above. For differences, the argument may be put in another way which may be more helpful. Our hypothesis is that the true value is zero, and the question we ask is, "What is the probability that random sampling will give a difference greater than that observed?" Now we do not know the sign, and it is purely an accident whether in comparing two means, $\bar{x}_a$ and $\bar{x}_b$ (say), we take $(\bar{x}_a - \bar{x}_b)$ or $(\bar{x}_b - \bar{x}_a)$, so we may always make the difference positive $(= + d$, say). Then in considering the sampling curve we must also do the same thing and only use the positive half of it, in which differences can only range from zero to plus infinity. Thus the probability that of all the *positive* differences that could be obtained from random sampling, one would be greater than $+ d$ is the ratio of the area beyond the ordinate at $+ d$ to the total area of the halved normal curve. In Fig. 3, OE is twice the standard error, and the area under the curve to the right of EF is 5 per cent. of the half of the curve to the right of the origin, O, or 2·5 per cent. of the full curve. Thus, there may be a distinction between the 5 per cent. *value* or *point* at which an ordinate cuts off a tail of 5 per cent. of the total area, and the 5 per cent. *level of significance*.

Sometimes, it is reasonable to regard a deviation corresponding to a single "tail" of 0·05 as lying on the 0·05 level of significance. For example, a strong believer in Scottish superiority might argue that Scotsmen cannot possibly be shorter than Englishmen, that they must either be equally tall or taller. He would only subject to statistical test a difference that showed Scotsmen as taller, and any difference of the opposite sign he would dismiss without test as being due to random errors. It is consistent with such an attitude to regard a difference of 1·65 times the standard error, corresponding to a single "tail" of 0·05 as lying on the 5 per cent. level of significance. It is only legitimate to do this, however, if the grounds for

deciding the sign of the difference or deviation (if real) are *a priori* and independent of the result given by the sample.

Groups of Samples

3.33. The probabilities deduced in accordance with the foregoing theory are only for single pairs of samples taken at random, and when there are several samples, the problem of significance is more complicated. If we had a hundred differences whose true value was zero, we should expect four or five of them to be greater than twice the standard error, but if we applied the simple theory by "rule-of-thumb," we should erroneously report them as real. Similar considerations must be applied to groups of less than one hundred tests and we will work out the significance of the biggest of n deviations.

TABLE 3.2

LARGEST $\frac{\text{DEVIATION}}{\text{STANDARD ERROR}}$ OF n LYING ON 0·05 LEVEL OF SIGNIFICANCE

n	$\frac{\text{Deviation}}{\text{Standard Error}}$
1	2·0
2	2·2
4	2·5
6	2·6
10	2·8

Suppose there are n independent differences between pairs of means; e.g. the aim may be to see if several treatments have an effect on some quantity and one of each pair of samples may be an untreated control (a separate one for each pair) and the other be given one of the treatments. We wish to test if the biggest ratio *difference/standard error of difference* is significant. Let the probability that errors of random sampling would give a *single* difference equal to or greater than d (say) be P (as found in the ordinary way from probability tables), and let P_n be the probability that the biggest of n would equal or exceed d. Then $(1 - P_n)$ is the probability that n differences would be less than d, and $(1 - P)$ that *one* random difference would be less; from Rule II (section 2.11) we have

$$(1 - P_n) = (1 - P)^n,$$

whence $$P_n = 1 - (1 - P)^n.$$

Table 3.2 has been worked out to show what value the largest of n deviations (in terms of the standard error) must reach to lie on the 0·05 level of significance for the normal distribution.

Sometimes we want to know if the difference between the largest and smallest in a group of sample means is significant; such a

TABLE 3.3

RANGE / STANDARD ERROR OF A DIFFERENCE ON 0·05 AND 0·01 LEVELS OF SIGNIFICANCE

Number of Samples	$P = 0·05$	$P = 0·01$
2	2·0	2·6
4	2·6	3·1
6	2·9	3·4
10	3·2	3·6

difference is a range. E. S. Pearson (1932) gives some values of the range at various levels of significance and Table 3.3 has been calculated from them. When there are four samples, a difference of 2·5 times is equivalent to one of twice the standard error for a pair.

We shall illustrate this by the data of Table 3.4, which have been obtained from Table 10.6 showing the mean corn yield per plot for a number of agricultural plots subjected to five treatments.

TABLE 3.4

Treatment	Mean Yield per Plot, Grammes
A	295·2
B	297·5
C	276·3
D	272·2
E	271·8

We are given that the standard error of any one mean is 8·712, so that the standard error of any random difference between two means is 12·32 grammes.* The largest difference is between treat-

* The treatments are "dummies." The standard error is deduced from Table 10.71.

ments B and E, and is 25·7 or 2·1 times the standard error. For a single randomly chosen pair of means, this difference would be significant, but here it is the largest difference in a set of five means, and from Table 3.3 we deduce that it is well below the 5 per cent. level of significance.

If it is desired to compare the means of several samples as a whole, there are other and more suitable methods that will be described in Chapter VI; but if the wish is to compare selected pairs, the above considerations show at least that more stringent tests of significance should be applied than when a single pair is taken at random.

Applicability of Normal Sampling Theory

3.4. The correspondence of a deviation of twice the standard error to the 0·05 level of significance depends on the sampling distribution of the mean being normal. Strictly, this only happens when the individuals in the sampled population are normal, but for large samples taken from most non-normal populations that are likely to be met with in practice, the distribution of the mean is nearly normal, and the normal sampling theory is sufficiently close an approximation for practical purposes. This frequently is a second reason for applying the methods of this chapter to large samples only.

The population and a sample may be characterised by constants other than the mean, and in fact any of the constants mentioned in Chapter I may be used. These all have their sampling distributions and, if they are known, deviations corresponding to various levels of significance may be calculated. For large samples, however, most of these distributions tend to be nearly normal, and as an approximation it is usual to apply the normal theory by calculating the standard error of the constant, and regarding deviations of twice this standard error as lying on the 5 per cent. level of significance. In this chapter, further applications of sampling theory will usually follow this procedure, but it is desirable for the sake of subsequent applications to discuss the use of skew sampling distributions. This discussion follows in the next section.

Tests of Significance Based on Skew Sampling Distributions

3.41. For a symmetrical distribution we decided in section 3.32 to regard the 2·5 per cent. points (on the positive and negative sides of the population value) as lying on the 5 per cent. level of

significance, and since the two deviations are equal, only one need be specified. When the distribution is not symmetrical, some modification is obviously necessary, and there are three possibilities.

(1) We may retain the equality of the positive and negative deviations, and regard as lying on the 5 per cent. level of significance a value such that the areas of the two tails beyond ordinates at these points together add up to 5 per cent. of the total area. In Fig. 6 (which is not drawn to any scale, and is only diagrammatic), O is the population value, and $OA = OA'$ is the significant deviation,

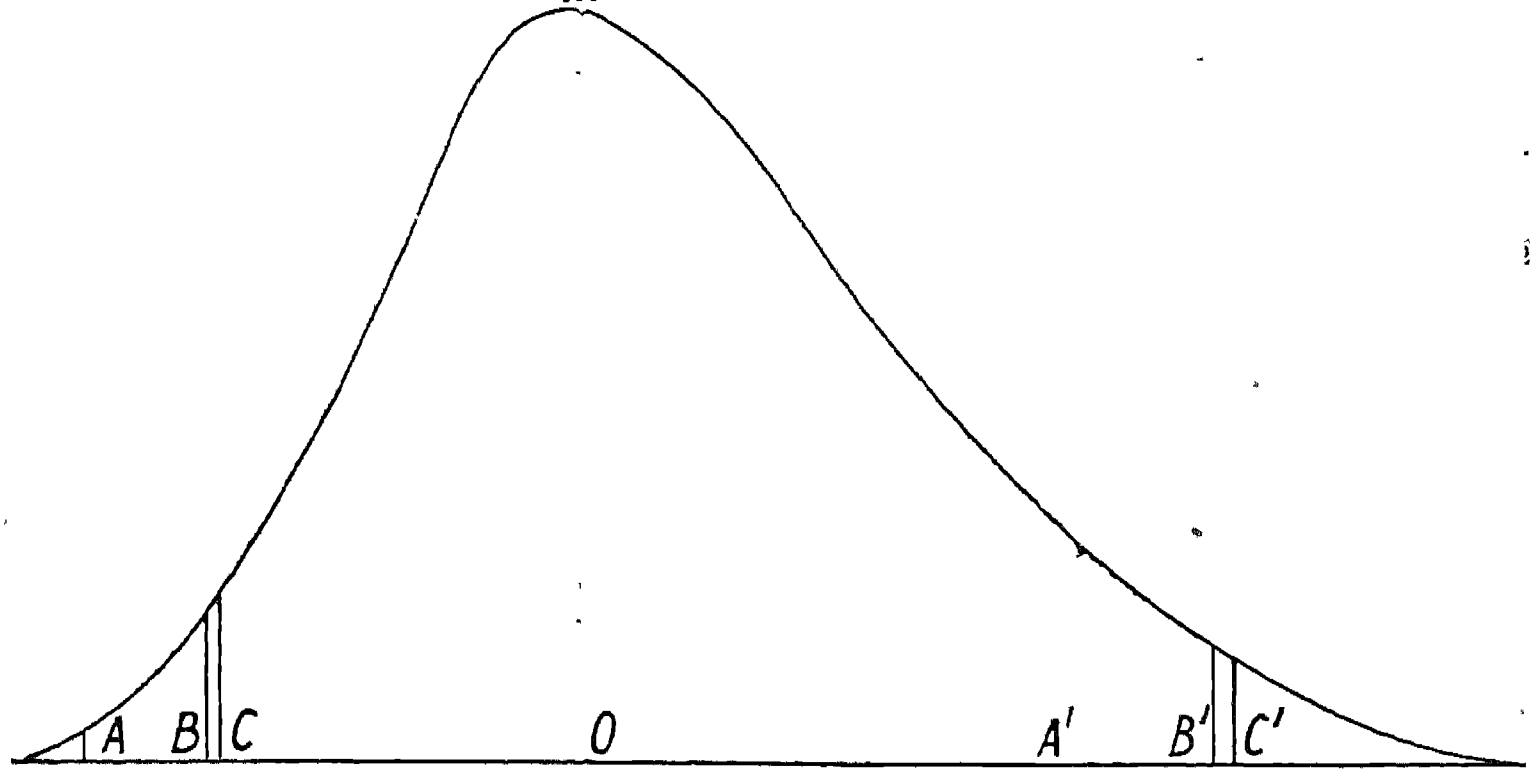


FIG. 6.

the areas beyond the ordinates at A and A' being 5 per cent. of the total.

Such a criterion is to be condemned, for it gives positive and negative deviations equal importance.

(2) Extending the argument already applied to differences in section 3.32, we may use separate tests for positive and negative deviations from the population value. If the deviation under test is positive, we use only that part of the sampling distribution on the positive side of the ordinate drawn at the population value, and regard as lying on the 5 per cent. level the deviation at which an ordinate cuts off a tail of 5 per cent. of the positive part of the curve. Similarly, if the deviation is negative we use only the negative part of the sampling distribution. In Fig. 6, OB and OB' are the two deviations, and the area to the left of the ordinate at B is 0·05 of that to the

left of OM, while the area to the right of the ordinate at B' is 0·05 of that to the right of OM. Because OM does not divide the total area into equal parts, the two tails are not equal and therefore are not 0·025 of the total area.

(3) We may regard as lying on the 5 per cent. level of significance the deviations beyond which the two tails are each equal to 2·5 per cent. of the total area under the curve. In Fig. 6, C and C' are about on these points. The shape of any frequency curve may be altered by plotting the abscissæ in terms of some function of x (e.g. putting $x' = \log x$), and it is theoretically possible to choose the function so that the curve reduces to the normal form. Such a procedure alters the relative positions of the mean and mode, but not of deviations defined by the ratios in which they divide up the area. Consequently 2·5 per cent. values of the transformed variate x' still remain 2·5 per cent. values of the untransformed variate x in the skew distribution. If the normal curve is regarded as the fundamental curve of errors, the deviations lying on the 5 per cent. level of significance according to this criterion would satisfy the normal criterion if the curve were transformed.

At the moment, both the second and third uses of skew curves for testing the significance of differences appear to be consistent extensions of the criteria developed for symmetrical curves, and it is difficult to decide between them. Fortunately, the difference in results obtained by the two methods is of no practical importance, and the last one will probably be most convenient in practice.

Sampling Errors of Various Constants

Means of Binomial and Poisson Distributions

3.51. When developing the binomial distribution in Chapter II we regarded the sets of trials as individuals and the number of occurrences per set as the variate. If, however, we regard the trials as individuals for which the variate can only take one of two characteristics, viz. success or failure, the set of n trials becomes a sample of n and the binomial distribution becomes a sampling distribution of the number of sucesses per sample. From this point of view, the small n of the binomial is equivalent to the large N of this chapter. As stated in section 2.5, when the number in the sample is large, the binomial distribution approaches the normal and the normal sampling theory may be applied.

For the binomial the standard error of the number of successes ($l = np$) is the standard deviation given in equation (2.2), p. 48, and is

$$\sqrt{npq} = \pm\sqrt{l - \frac{l^2}{n}};$$

on dividing thus by n we obtain,

$$\textit{Standard Error of } p = \pm\sqrt{\frac{p - p^2}{n}}.$$

Darbishire (1904), on crossing waltzing with normal mice, found in the F_2 generation 458 normals and 97 waltzers (total 555). The

TABLE 3.5

Diet	Males	Females	Total Young	Percentage Males
Vitamin B Deficient ..	123	153	276	44·57
Vitamin B Sufficient ..	145	150	295	49·15
Totals	268	303	571	—

Mendelian expectation for l is 416 normals with a standard error of

$$\pm\sqrt{416 - \frac{416^2}{555}} = \pm 10{\cdot}2.$$

The difference between the actual and expected number of normal mice is 42, and being 4·1 times the standard error is significant; it could only arise from random sampling four times in 100 000 trials.

Parkes and Drummond (1925) give the data of Table 3.5, showing the effect of vitamin B deficiency on the sex-ratio of the offspring of rats.

In comparing the sex-ratios, we are not comparing an experimental ratio with a theoretical one, but two experimental ones. Hence, we do not know the values of p in the infinite population required for the calculation of the standard errors; we will use the sample values as an approximation. The standard errors of p are

$$\pm\sqrt{\frac{0{\cdot}445\,7 - 0{\cdot}198\,6}{276}} = \pm 0{\cdot}029\,9$$

and

$$\pm\sqrt{\frac{0{\cdot}491\,5 - 0{\cdot}241\,6}{295}} = \pm\,0{\cdot}029\,1,$$

and that of the difference is $\pm\,0{\cdot}041\,7$, or $\pm\,4{\cdot}17$ per cent. The difference in sex-ratio is $4{\cdot}58$, and being only $1{\cdot}10$ times its standard error is insignificant; it would arise from random errors nearly 27 times in 100 trials.

3.52. When the mean number of failures (m) per set is large, the Poisson distribution also approaches the normal form. Then the standard error per set is $\sqrt{m}$, the standard error of the mean of N sets is $\sqrt{m/N}$, and that of the total number of failures in N sets, mN is

$$N\sqrt{\frac{m}{N}} = \sqrt{mN}.$$

Thus for the distribution of yeast cells of Table 2.4, the total number of cells counted is 1 872, with a standard error

$$\pm\sqrt{1\,872} = \pm\,43{\cdot}3;$$

the mean number of cells per square is $4{\cdot}68$, and its standard error is

$$\pm\sqrt{\frac{4{\cdot}68}{400}} = \pm\,0{\cdot}108.$$

Measures of Dispersion

3.53. The standard deviation estimated from the variance has a standard error of

$$\frac{\sigma}{\sqrt{2N}}$$

where σ is the standard deviation of the population and N is the size of the sample. Its distribution is not normal, but for large samples it approaches normality; e.g. at $N = 100$, $\beta_1 = 0{\cdot}005\,1$ and $\beta_2 = 3{\cdot}000\,0$. Let us see if the Scotsmen of Table 3.1 are really more regular in height, as well as being taller on the average, than the Englishmen. The standard error of a standard deviation is

$1/\sqrt{2}$ times that of a mean, so for the difference of standard deviations of the table it is

$$\frac{1}{\sqrt{2}} \times 0{\cdot}075\,9 = \pm 0{\cdot}053\,7.$$

The difference ($2{\cdot}548 - 2{\cdot}480 = 0{\cdot}068$) being only $1{\cdot}27$ times its standard error is not significant; in fact, random sampling would give as big a difference about one trial in five ($P = 0{\cdot}2$).

Only when found from the second moment has the standard deviation the above standard error. We shall now deal with its standard error when estimated from the range. E. S. Pearson (1926 and 1932) has worked out the sampling distribution of the range fairly fully, and although for no size of sub-sample is it normal, it is only moderately skew for those of about 10; for sub-samples of 10, $\beta_1 = +0{\cdot}156$ and $\beta_2 = 3{\cdot}22$. In such circumstances, we may use the standard error as an approximate description of the sampling errors. Pearson gives the standard error or deviation of the range for several sub-samples, and for those of 10 it is $0{\cdot}797\,\sigma$, where σ is the true standard deviation. Suppose the sample of N contains m sub-samples of 10, so that $N = 10\,m$. Then from section 2·71,

$$\textit{Standard Error of Mean Range} = \frac{0{\cdot}797\,\sigma}{\sqrt{m}}.$$

We also have from Table 1.2,

$$\textit{Estimated Standard Deviation} = \frac{\textit{Mean Range}}{3{\cdot}078},$$

whence

$$\textit{Standard Error of Estimated Standard Deviation} = \frac{\textit{Standard Error of Mean Range}}{3{\cdot}078} = \frac{0{\cdot}797\,\sigma}{3{\cdot}078\sqrt{m}} = \frac{1{\cdot}158\,\sigma}{\sqrt{2N}}.$$

This may be compared with the standard error of the estimate obtained from the variance.

In large samples, the standard error of the mean deviation is

$$1{\cdot}068\,\frac{\delta}{\sqrt{2N}} = 0{\cdot}861\,\frac{\sigma}{\sqrt{2N}},$$

where δ is the population value of the mean deviation. It follows that the standard error of the standard deviation estimated from the mean deviation is

$$1{\cdot}253 \times 0{\cdot}861 \frac{\sigma}{\sqrt{2N}} = 1{\cdot}068 \frac{\sigma}{\sqrt{2N}}.$$

Constants of Shape

3.54. The β ratios are useful for testing the departure of any data from normality, but their distributions in samples drawn from an infinite population of that form have not yet been worked out. The first four moments of the sampling distributions have been derived by Fisher (1930*a*), and from approximate values E. S. Pearson (1930) has determined appropriate empirical frequency curves of K. Pearson's system. Curves found in such a way usually give good approximations to the actual distributions.

To test asymmetry, $\gamma_1 = \sqrt{\beta_1}$ (with the same sign as μ_3) is used. For samples of N from the normal population, this has a standard error of $\sqrt{6/N}$ (to a first approximation), and the distribution itself is so nearly normal that a deviation of twice the standard error lies practically on the 0·05 level of significance.

The standard error of β_2 is $\sqrt{24/n}$ (to a first approximation), but the distribution is so skew, and this approximation is so poor, that it does not give a reliable test. Pearson's tables (*loc. cit.*) give for samples from the normal population, values of β_2 lying on each side of the population value (3·0) which cut off tails of 5 and 1 per cent. of the whole curve; these tables should be used. From the discussion in section 3.41, it is suggested that the 5 per cent. level of significance should be taken as lying between the 5 and 1 per cent. values.

For the distribution of heights of Table 1.5, we have

$$\gamma_1 = -0{\cdot}116 \pm 0{\cdot}074\,6, \quad \beta_2 = 2{\cdot}908 \quad \text{and} \quad n = 1\,078.$$

γ_1 is less than twice its standard error, and so is insignificant, while from Pearson's table, when $n = 1\,000$, the lower value of β_2 with a "tail" of 0·05 is 2·76, and since the value above is nearer 3 than that, we may conclude that the data are, as far as we can tell, from a normal population.

Standard Errors of Functions of Statistical Constants

3.55. Sometimes it is desired to calculate the standard errors of functions of one or more statistical constants, the errors of the individual constants being known. We have already had an example in the difference between two means; the ratio and product of two means, and the square of the standard deviation are other examples. We shall describe here an approximate method that is applicable to large samples. Where there are more than one constant we shall assume they are obtained from independent samples except in one instance.

Let the population values of the constants be α and β, let the function be

$$\lambda = f(\alpha, \beta)$$

and let the corresponding sample values for any one pair of samples be l, a and b, where

$$l = \lambda + \delta\lambda, \quad a = \alpha + \delta\alpha \quad \text{and} \quad b = \beta + \delta\beta,$$

δ is used as a sign meaning a deviation. Then for the pair of samples

$$l = f(\alpha + \delta\alpha, \beta + \delta\beta).$$

It can be shown that, for the purposes of deducing standard errors and variances, the relations between $\delta\lambda$, $\delta\alpha$ and $\delta\beta$ may be approximately derived by regarding them as mathematical differentials, so that

$$\delta\lambda = \frac{\partial f}{\partial \alpha}\delta\alpha + \frac{\partial f}{\partial \beta}\delta\beta$$

The degree of approximation involves neglecting quantities of the order $1/N$ compared with unity, where N is the number in the sample.

The mean values of the squares of $\delta\lambda$, $\delta\alpha$ and $\delta\beta$ for all possible pairs of samples from the population are the squares of the corresponding standard errors and may be written $(SE_l)^2$, $(SE_a)^2$ and $(SE_b)^2$. Hence, by squaring the terms of the above equation and finding the means for all pairs of samples,

$$(SE_l)^2 = \left(\frac{\partial f}{\partial \alpha}\right)^2 (SE_a)^2 + \left(\frac{\partial f}{\partial \beta}\right)^2 (SE_b)^2 + 2\frac{\partial f}{\partial \alpha}\frac{\partial f}{\partial \beta}[\delta\alpha\ \delta\beta]$$

where $[\delta\alpha\ \delta\beta]$ is the mean value of the product of the deviations $\delta\alpha$ and $\delta\beta$ for all pairs of samples. If the two samples in the pair

are independent, this mean product is zero, as will be shown in Chapter VII.* Hence, for independent pairs of samples,

$$(SE_l)^2 = \left(\frac{\partial f}{\partial \alpha}\right)^2 (SE_a)^2 + \left(\frac{\partial f}{\partial \beta}\right)^2 (SE_b)^2 \quad . \quad . \quad (3.6)$$

This equation may easily be extended to three or more statistical constants.

When the function is the difference between two means, equation (3.1), section 3.21, follows from (3.6) directly.

By way of example, let us deduce the standard error of the variance v of a sample in terms of that of the standard deviation s, so that $l = v$, $a = s$ and there is no b. Let the population value of the standard deviation be σ; then

$$v = s^2$$

$$(SE_v)^2 = 4\sigma^2(SE_s)^2 = \frac{2\sigma^4}{N}.$$

As another example, let l be the ratio between two means $\bar{x}$ and $\bar{y}$, with corresponding population values of λ, $\bar{\xi}$ and $\bar{\eta}$. Then

$$l = \frac{\bar{x}}{\bar{y}},$$

$$(SE_l)^2 = \lambda^2\left\{\frac{(SE_x)^2}{\bar{\xi}^2} + \frac{(SE_{\bar{y}})^2}{\bar{\eta}^2}\right\}.$$

The standard errors of $\bar{x}$ and $\bar{y}$ have already been given in section 2.71.

It has been shown in section 2.72 that the mean and standard deviation estimated from the same sample are independent, and it follows from this that equation (3.6) may be used to determine the standard error of the coefficient of variation. If 100 κ, $\bar{\xi}$ and σ are respectively the population values of the coefficient of variation, mean and standard deviation, corresponding to λ, α and β, and k is the sample estimate of κ, it is easy to see from equation (3.6) and the standard errors given in sections 2.71 and 3.53 that the standard error of the coefficient of variation is

$$100\,\kappa\sqrt{\frac{2\kappa^2 + 1}{2N}}.$$

* After reading Chapter VII, readers will be able to evaluate the product term and so deal with constants that are not independent.

Determination of Population Value from Sample

3.6. We have seen how a statistical constant determined from a sample may be used to test some hypothesis concerning the population value. Sometimes, however, we are unable to postulate the population value, of which we look to the sample to teach us something. This process of estimating the population value from the sample or of arguing from the particular to the general is the

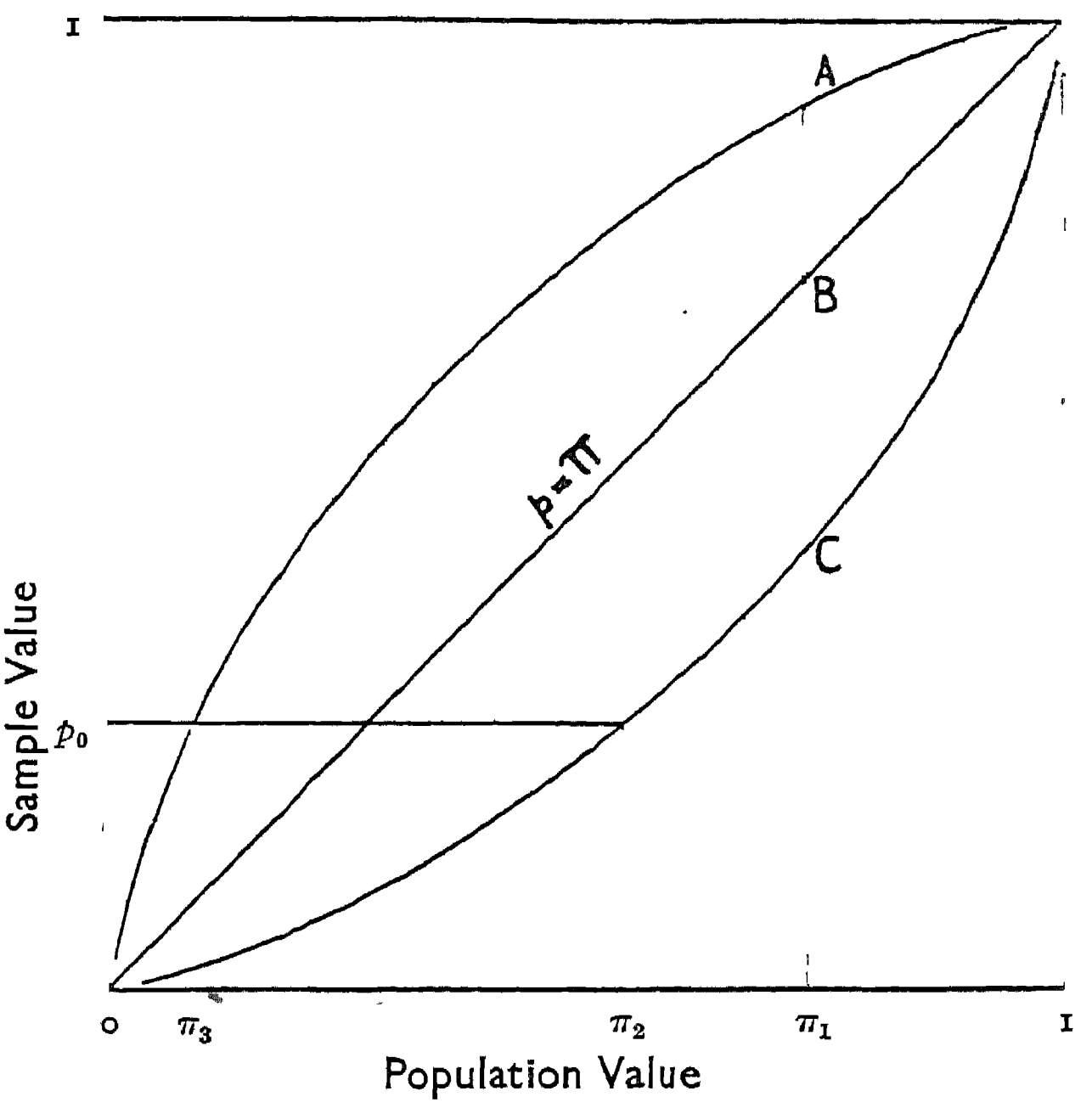


Fig. 7.

inductive method, and the question arises as to the accuracy of the estimate given by the sample. This may be answered if the sampling distribution of the constant is known, and we shall now illustrate the kind of answer provided by referring to samples of 100 individuals taken to measure the proportion having a given character, i.e. the proportion of successes.

The population value of this proportion may be denoted by π* and any sample value by p. Then a diagram like that of Fig. 7 may be drawn (this is not drawn to scale), where for any population

* Here π is not 3·14159 . . .

value π_1 the point B for which $p = \pi_1$ represents the mean of all the sample values, and the points A and C represent values of p lying on the 5 per cent. level of significance. These points may be found from the sampling distribution of p, which is given by the terms of the binomial $(\pi_1 + \overline{1 - \pi_1})^{100}$, and if we may assume this to be approximately normal,

$$AB = BC = \textit{Twice the Standard Error} = 2\sqrt{\pi_1(1 - \pi_1) \div 100}.$$

For example, if $\pi_1 = 0{\cdot}8$, B is at $p = 0{\cdot}8$, and the standard error is $\sqrt{0{\cdot}8 \times 0{\cdot}2 \div 100} = 0{\cdot}04$; hence A is at $p = 0{\cdot}8 + 0{\cdot}08 = 0{\cdot}88$ and C is at $p = 0{\cdot}8 - 0{\cdot}08 = 0{\cdot}72$. In this way, points corresponding to A and C may be found for all possible values of π, and these lie on the curved lines shown diagrammatically in Fig. 7. The points corresponding to B lie on the line $p = \pi$. Then we know that for any one value of π, 95 per cent. of the sample values p lie within the limits represented by the points on the curved lines at which an ordinate drawn at π cuts them, and so by adding the experience for all populations, we know that 95 per cent. of the sample values lie within the limits of the curved lines, whatever may be the relative frequencies with which the different values of π occur.

If in making inferences from samples it is assumed that all sample-population points lie within the limits, i.e. that for every sample value p_0 (say) the population value lies between π_3 and π_2, we shall be right in 95 per cent. of the inferences. In this way, it is possible to make an estimate from a sample of the range within which the population value lies, with a given long-run risk of being wrong.

Fisher (1930*b*) has termed the limits represented by the curved lines the *fiducial limits* and the corresponding proportion of correct inferences (95 per cent. in Fig. 7) the *fiducial probability*. Neyman (1934) uses the terms *confidence limits* and *confidence coefficient*. It is possible, of course, to determine limits corresponding to other fiducial probabilities, e.g. for the 99 per cent. level, and for any statistical constant that has a known continuous sampling distribution. In order to determine the sampling distribution of the constant, it is usually necessary to assume the form of the population, e.g. normal, and it may be necessary to assume the standard deviation. When the distribution is skew, difficulties of the kind mentioned in section 3.41 have to be faced.

In deriving the limits in Fig. 7 by the methods outlined, the

result would be limits of p for known values of π; for practical purposes it is more convenient to have limits of π for known equidistant values of p. It is not proposed here to go into the method of arriving at this desired form of expression (simple inverse interpolation is effective), but it is pointed out that the limits of π corresponding to a sample value p are *not* $p \pm 2\sqrt{p(1-p) \div 100}$. For example, when $p_0 = 0{\cdot}2$, $\pi_2 = 0{\cdot}291$ and $\pi_3 = 0{\cdot}132$. It will be noted that the limits of π are not even equidistant from p. These conditions arise when the standard error of a statistical constant depends on the population value of the constant as for the mean of a binomial or Poisson series. When this is not so, and the standard error is independent of the constant, the 95 per cent. fiducial limits of the population value are two parallel straight lines, and for any one sample value are at twice the standard error above and below that value, assuming a normal sampling distribution. Thus the 95 per cent. limits of the mean height of Englishmen in Table 3.1 are $67{\cdot}437\,5 \pm 2 \times 0{\cdot}032\,38$, i.e. $67{\cdot}502\,3$ and $67{\cdot}372\,7$ inches.

Fiducial probability is not the same as *inverse probability*, although is appears superficially to be so. An inverse probability statement would be applied to any particular instance and would be of the form: "given a sample value p_0, the probability of the population value lying between π_2 and π_3 is $0{\cdot}95$" (see Fig. 7). This involves the conception of a super-population of population values corresponding to the one sample value p_0, 95 per cent. of which are between π_2 and π_3. Such a conception lies behind any statistical interpretation of inverse probability and presents many difficulties, one of the chief of which is that there is no basis for making any assumption regarding the distribution in the super-population. The fiducial probability statement says that for all possible sample values taken from any populations, 95 per cent. of the population values of π lie between the limits shown in Fig. 7; no assumption is made as to the distribution of the population values.

Fiducial limits can only be calculated in the way shown in so far as the sampling distribution is continuous or may be approximately represented by a continuous distribution.

Choice of Statistical Constants

3.7. We have seen in Chapter I that there may be several statistical constants for measuring one characteristic of a sample or popula-

tion; for example, the standard deviation, mean deviation, quartile deviation and mean range are all measures of dispersion. The question arises: Are there any grounds on which one constant is to be preferred to another? Fisher's theory of estimation (1922, 1925*a* and 1936) provides an answer, and we shall give as much of it as seems necessary to show the kind of answer provided.

So far, we have regarded statistical constants as convenient measures of various characteristics of samples and populations, and have assumed some mathematical form of distribution for the population more or less incidentally. In the theory of estimation it is necessary to assume or specify some mathematical form of distribution in the population as a starting-point. The equation of this form has one or more constants or parameters that may vary from one population of that form to another, and the particular values of which define any given population. For example, if the assumed form of the population is the normal curve described by equation (2.4), the two parameters are m and σ, and given values of these define a particular normal population. On this view, m and σ are not thought of as the mean and standard deviation, i.e. as measures of position and dispersion; they are thought of as mathematical parameters. Similarly, statistical constants (Fisher calls them *statistics*) calculated from samples are not descriptive averages; they are estimates of the parameters in the equation, and it is on this basis that the relative values of equivalent statistics are judged. Here, we need concern ourselves with two only of the criteria given by Fisher; they are consistency and efficiency.

Consistency.—A consistent estimate of a parameter must equal the parameter when it is derived from the whole distribution in the population. In a normal population, all the above measures are consistent estimates of σ. This criterion is an obvious one and is satisfied by all statistical estimates in common use.

Efficiency.—This criterion applies to those statistical estimates from large samples for which the sampling distribution is known and is normal, so that it is completely described by the mean and standard error.* The mean of the estimate in the sampling distribution may differ from the parameter, but this difference may be calculated and allowed for as bias. The standard error defines the

* As an approximation, the ideas and criteria may be applied where the sampling distribution is nearly normal, i.e. to all constants the standard errors of which are given in this chapter, except β_2.

inaccuracy of the estimate; here it will be more convenient to deal with the square of the standard error, viz. with the *variance* of the estimate.

For statistics of the class with which we are now dealing the variance in samples of N may be expressed in the form

$$\frac{1}{IN}$$

where I depends on the statistic and is called its *intrinsic accuracy*. For example, the intrinsic accuracy of the mean as an estimate of m in the normal population is $1/\sigma^2$ and that of the square root of the second moment as an estimate of σ is $2/\sigma^2$.

We have seen in section 3.53 that the various estimates of the standard deviation have differing standard errors, and hence intrinsic accuracies. Moreover, an increase in I has the same influence on the sampling variance as a proportionate increase in N, so that differences between intrinsic accuracies may be expressed in terms of N. Thus, for σ estimated from the second moment, $I = 2/\sigma^2$, and for the estimate obtained from the mean range in sub-samples of 10, $I = 2/1{\cdot}34\,\sigma^2$. The variance of the estimate from the range in a sample of 100 is equal to that of the estimate from the second moment in a sample of $100/1{\cdot}34 = 75$, for

$$\frac{\sigma^2}{2 \times 75} = \frac{1{\cdot}34\,\sigma^2}{2 \times 100}.$$

The loss in accuracy due to using the range instead of the second moment is equivalent to a 25 per cent. reduction in the size of the sample. We may say that the estimate from the range has an efficiency of 75 per cent. of the estimate from the second moment.

For each parameter, there is a class of statistics that have the same intrinsic accuracy, that is greater than the intrinsic accuracy of all other estimates, and these are called *efficient* statistics. The ratio of the intrinsic accuracy of any given statistic to that of an efficient one is called the *efficiency* of the given statistic. The mean and standard deviation as defined in sections 1.22 and 1.23 are efficient estimates of m and σ in the normal population.

When using one estimate of a parameter, we may regard the sample as giving a certain quantity of information regarding the parameter, that increases in proportion to the size of the sample.

Each individual may be regarded as contributing a unit amount of information, and the connection between the intrinsic accuracy and size of sample suggests that the intrinsic accuracy should be the unit. Fisher in his later writings (1935) has referred to intrinsic accuracy as the *quantity of information given by a single observation*. In these terms, the mean range in sub-samples of 10 may be said to give three-quarters of the amount of information given by the second moment regarding the parameter σ. Just as for a given estimate the number of individuals is additive in the sense that two independent samples of N_1 and N_2 give the same amount of information as one sample of $N_1 + N_2$, so for different estimates and samples, independent quantities of information are additive.

When combining estimates of a parameter that have normal sampling distributions, the estimates being from different samples, we may take a weighted mean, using the quantities of information as weights. This is analogous to combining estimates of the mean, say, by using the numbers in the samples as weights. The variance of the combined estimate is the inverse of the quantity of information. This result is used in sections 8.22 and 11.73.

There are several reasons why efficient statistics should be used except in special circumstances. The use of an inefficient statistic is equivalent to throwing away part of the data, and such a course is usually only justifiable if there is a proportionate saving in labour, say in computation. When efficient statistics are used, mistaken inferences of the second kind mentioned in section 3.3 are reduced to a minimum. Further, populations are estimated with maximum precision by the use of efficient statistics, for these have fiducial limits corresponding to a given fiducial probability, that are closer than for any other estimates of the same parameter. The methods of the next chapter apply only when efficient estimates are used.

A distinction must here be made between different statistical constants that are merely mathematical transformations of one constant, and constants that are essentially different estimates and are only related statistically. For example, the variance and standard deviation are mathematically equivalent; samples that have the same variance must necessarily have the same standard deviation and the two constants are equally efficient. On the other hand, the standard deviation, mean deviation, and mean range in sets of 10 are only statistically related in that the relations given in section 1.23 are true only *on the average* and do not necessarily apply to any

particular sample; different samples having the same standard deviation may have different mean deviations and mean ranges. These estimates do not have the same efficiency.

Method of Maximum Likelihood

3.8. This is a general method for arriving at estimates of parameters of distributions, and Fisher has shown that estimates given by this method are efficient, provided efficient statistics of the parameter exist.

We have already seen in section 2.72 how the probability of a given sample may be deduced when the population and its parameters are known. When the form of population is known the expression for that probability (without the differential terms) for any assumed values of the parameters is termed the *likelihood* of the assumed values. The particular values that make this likelihood a maximum are the *maximum likelihood* estimates. They are obtained by differentiating the logarithm of the likelihood with respect to the parameters, equating to zero and solving the equations.

For the normal distribution with assumed parameters $\bar{\xi} = \hat{x}$ and $\sigma = \hat{s}^*$ and a sample of N, the logarithm of the likelihood (to the base e) is:

$$L = -\frac{N}{2}\log 2\pi - N\log \hat{s} - \frac{1}{2}\left\{\frac{(x_1-\hat{x})^2}{\hat{s}^2} + \frac{(x_2-\hat{x})^2}{\hat{s}^2} + \ldots \frac{(x_N-\hat{x})^2}{\hat{s}^2}\right\} \quad . \quad (3{\cdot}7)$$

If this is differentiated with respect to $\hat{x}$ and the differential equated to zero, it is easily deduced that

$$\hat{x} = \frac{x_1 + x_2 + \ldots x_N}{N}.$$

Thus, the mean of the sample, as defined in section 1.22, is the maximum likelihood estimate of $\bar{\xi}$. Similarly, by differentiating L in (3.7) with respect to $\hat{s}$ and equating to zero, the maximum likelihood estimate of σ is found to be the standard deviation as defined in section 1.23.

* The Latin letters denote estimates from the sample and the sign $\wedge$ denotes the particular kind of estimate dealt with in this section.

3.81. When the data are in the form of a frequency distribution with frequencies of $n_1\, n_2 \ldots n_s \ldots$ in the groups, the logarithm of the likelihood is

$$L = \underset{s}{S} n_s \log \hat{n}_s \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.8)$$

where $\underset{s}{S}$ is the summation over all groups and $\hat{n}_s$ is the estimated frequency in the sth group, determined from the equation for the assumed population and expressed in terms of the unknown parameters.

For example, for the Poisson distribution the proportion of frequency with s successes is

$$\frac{\mu^s}{s!} e^{-\mu}.$$

If $\hat{m}$ is the maximum likelihood estimate of μ, and the total number in the sample is N,

$$L = \underset{s}{S} n_s \left\{ s \log \hat{m} - \hat{m} - \log \frac{s!}{N} \right\}.$$

By equating to zero the first differential of this with respect to $\hat{m}$ we have

$$\hat{m} = \frac{\underset{s}{S} s n_s}{\underset{s}{S} n_s} = \frac{\underset{s}{S} s n_s}{N}.$$

Thus, the efficient estimate of μ is the mean of the distribution. This result is not without interest, since we have seen in section 2.3 that the second moment of the distribution is also a possible estimate of μ, and hitherto no rational grounds have been advanced for regarding it as an inferior estimate.

The method of maximum likelihood may be used whenever a distribution can be expressed in terms of parameters that are required to be estimated from the sample. Fisher (1936*a*) has used it in genetical studies for estimating linkages.

3.82. When an efficient statistic has a normal sampling distribution, the square of the standard error or variance in a large sample may be obtained from the second differential of the logarithm of the likelihood L. Generally, if κ is the parameter, $\hat{k}$ the maximum likelihood estimate, and $\sigma_{\hat{k}}$ is the standard error of $\hat{k}$,

$$-\frac{1}{\sigma_{\hat{k}}^2} = \left[\frac{\partial^2 L}{\partial \hat{k}^2} \right] \quad . \quad . \quad . \quad . \quad . \quad (3.9)$$

where the square brackets [ ] signify "the mean for all possible sample values of the term contained therein."

For the normal distribution, on differentiating equation (3.7) twice, we find

$$\frac{\partial^2 L}{\partial \hat{x}^2} = -\frac{N}{\hat{s}^2}$$

and the mean value of $\hat{s}^2$ for all possible samples is the population value σ^2, whence

$$\sigma_{\hat{x}}{}^2 = \frac{\sigma^2}{N}.$$

This is an alternative derivation of the square of the standard error of the mean.

CHAPTER IV

GOODNESS OF FIT AND CONTINGENCY TABLES

4.1. So far we have been using as expressions of differences between distributions, constants like the mean, standard deviation, β_1 and β_2, which summarise their chief properties or are estimates of their parameters. Except for populations of particular forms, these do not express all the features of the distributions. It is desirable to have some index which measures the degrees of difference between the actual frequencies in the groups, and so compares all the essential features. Such is K. Pearson's (1900) χ^2, which we shall first use to measure the deviations of an experimental distribution from the form of some hypothetical population. Both must be grouped in the same way, and the theoretical distribution must be adjusted to give the same total frequency; then if ν_s is the number of observations in any one group in the theoretical distribution, and n_s is the corresponding number in the experimental one,

$$\chi^2 = \underset{s}{S}\frac{(n_s - \nu_s)^2}{\nu_s} \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.1)$$

where $\underset{s}{S}$ is the summation over all groups. It will be appreciated that since $(n_s - \nu_s)$ is squared, all differences in frequency, whether positive or negative, add a positive amount to χ^2, and further that the greater these differences are, the greater is χ^2; if the two distributions are exactly alike, χ^2 is zero.

Sampling Distribution of χ^2

4.2. In using χ^2 to test whether one distribution differs from the other, we must remember that because of random errors, χ^2 will never (or hardly ever) be zero, and we need to know its sampling distribution so that we can tell the probability of an observed χ^2 being given by a random sample from the hypothetical population. As usual, if that probability (which is symbolised by P) is low enough, the χ^2 is said to be significant, and it is unreasonable to suppose that such a significant value could be a result of sampling errors alone. Subject to certain restrictions, the distribution of χ^2 is given by the equation,

$$df = k\chi^{g-1}e^{-\frac{1}{2}\chi^2}d\chi, \quad . \quad . \quad . \quad . \quad (4.2)$$

where df is the element of frequency between ordinates drawn at χ and $\chi + d\chi$, the frequency between $\chi = 0$ and $\chi = \chi_1$ being $\int_0^{\chi_1} df$, k is a constant and g is the number of *degrees of freedom*.

This conception of *degrees of freedom* is not altogether easy to attain, and we cannot attempt a full justification of it here; but we shall show its reasonableness and shall illustrate it, hoping that as a result of familiarity with its use the reader will appreciate it. Here the number of degrees of freedom is the number of groups, modified. Clearly, since the error in each group in the experimental distribution contributes a positive amount to χ^2, the greater the number of groups the larger would χ^2 be expected to be, as a result of random variations alone, and account of this is taken through the quantity g. Further, it is often the practice to fit the theoretical distribution to the observations by calculating constants from the sample, just as in the example of section 2.51 we fitted a normal curve by making its mean, standard deviation and total equal to those of the sample. If we wish to test the adequacy of the theoretical form, further account must be taken of the degree to which we have made it fit the observations by this method. Suppose, in an extreme case, there were g' groups and we fitted a curve involving g' constants which were calculated from the data; then the two distributions would agree exactly and χ^2 would be zero because sampling errors would have had no play. To take account of this second factor, we must subtract from the number of groups (say g') the number of constants that have been determined from the data in fitting,* in order to obtain the degrees of freedom (g). Every constant so determined has the effect, from the point of view of χ^2, of reducing the number of groups by one, and the number of degrees of freedom may be regarded effectively as the number of independent groups remaining to contribute to χ^2. When only the totals have been made equal, $g = g' - 1$, but in a case like the fitting of a binomial to the number of germinating seeds in Table 2.3, the theoretical distribution has been adjusted to make its mean and total both equal to the sample, and $g = g' - 2$.

One restriction under which equation (4.2) for the sampling distribution of χ^2 applies is that no group contains very few individuals; 10 is about the lower limit. To arrive at the distribution

* Provided these constants are determined by processes which satisfy the criterion of *efficiency*.

of χ^2 experimentally, we should have to take many samples. Now the probability of any random individual belonging to the sth group (say) is ν_s/N, where N is the total number in the sample, and consequently in all the samples, the actual numbers falling into the sth group (n_s) will vary according to the binomial

$$\left[\frac{\nu_s}{N}+\left(1-\frac{\nu_s}{N}\right)\right]^N$$

(this is a direct application of the binomial theory given in section 2.2). If ν_s is not too small, this binomial approximates to the normal distribution, and it is on the assumption that n_s is distributed normally that equation (4.2) is obtained. This is the reason why in using χ^2 no group should contain much fewer than ten individuals, and to satisfy this in practice the tail groups, which often have very low frequencies, are combined.

A second restriction is the usual one that all the individuals in the sample must be independent. Indeed, χ^2 may be used as a test of independence.

In its sampling distribution, χ^2 may vary between zero and plus infinity, and since there is no question of negative deviations, and we are asking if the observed χ^2 is really greater than zero, the value at which an ordinate cuts off a tail of 5 per cent. also lies on the 5 per cent. level of significance. Since, in general practice, we should not test a very small value of χ^2 for the significance of its deviation from the average, we do not include the tail near $\chi^2 = 0$ in the test for significance; this is in accordance with the principles discussed in section 3.32.

There are two sets of tables of the probability integral of equation (4.2). The first, calculated by Elderton, is included in Pearson's *Tables for Statisticians and Biometricians*, and gives the areas of the tails beyond integral values of χ^2 for different values of n', where n' is one more than the number of degrees of freedom ($= g + 1$). This was originally calculated for the case where the theoretical distribution is only made to agree with the data in respect of its total, and was wrongly applied to other cases for some time before the necessity for correcting to degrees of freedom (which was pointed out by Fisher in 1922*b*) was realised. The other tables by Fisher (1936*a*) give values of χ^2 lying on different levels of significance, and in these the degrees of freedom are used directly and are called n.

The χ^2 test is more universal in its application than most others

in that there is no assumption made as to the normality of the distributions being compared.

First let us test the fit of the normal curve to the distribution of heights of men in Table 1.5; the arithmetical operations are set out in Table 4.1, and it will be noticed that the tail groups have been lumped together, giving 14 groups. Three constants have been fitted (total, mean and standard deviation), leaving 11 degrees of

TABLE 4.1

Stature in Inches	Frequencies: Observed (n_s)	Frequencies: Expected (ν_s)	$(n_s - \nu_s)$	$\frac{(n_s - \nu_s)}{\nu_s}$	$\frac{(n_s - \nu_s)^2}{\nu_s}$
below 61·5	14·5	11·8	2·7	0·229	0·62
61·5–	17	17·7	− 0·7	− 0·040	0·03
62·5–	33·5	35·5	− 2·0	− 0·056	0·11
63·5–	61·5	62·8	− 1·3	− 0·021	0·03
64·5–	95·5	96·7	− 1·2	− 0·012	0·01
65·5–	142	130·1	11·9	0·091	1·08
66·5–	137·5	153·0	− 15·5	− 0·101	1·57
67·5–	154	157·1	− 3·1	− 0·020	0·06
68·5–	141·5	141·0	0·5	0·004	0·00
69·5–	116	110·5	5·5	0·050	0·27
70·5–	78	75·7	2·3	0·030	0·07
71·5–	49	45·2	3·8	0·084	0·32
72·5–	28·5	23·7	4·8	0·203	0·97
over 73·5	9·5	17·2	− 7·7	− 0·448	3·45
Total ..	1 078	1 078·0	0·0	—	$\chi^2 = 8·59$ $n' = 12$ $P = 0·66$

freedom, and we wish to see if the χ^2 (8·59) resulting from those 11 can be attributed to random variations. Entering Elderton's table at $n' = 12$, we find $P = 0·66$, and this is so large that we might reasonably suppose the deviations to have arisen from errors of random sampling, and we say that the normal curve gives a good fit. Indeed, 66 random samples in 100 would have given a χ^2 equal to or greater than 8·59. In computing χ^2 the total of the fourth column (= 0) checks the accuracy of the subtractions; and the terms in the sixth column are the products of those in the fourth and fifth.

Similarly, the distribution of germinating seed in Table 2.3 gives a χ^2 of 2·897 when the groups containing four or more seeds are combined; there are five groups and three degrees of freedom (the mean and total have been equalised), and entering Elderton's table at $n' = 4$, we find $P = 0{\cdot}41$, showing that the fit is a good one.

An experimental distribution of flower colours and the Mendelian expectations are shown in Table 4.2 (Wheldale, 1907). There are 7 groups, and as only the total has been found from the sample (the expected frequencies, being determined from Mendelian

TABLE 4.2

Colour of Flowers	Expected Frequency	Experimental Frequency
Magenta	118·34	107
Magenta delila	39·44	42
Ivory	52·59	67
Crimson	39·44	42
Yellow	17·53	24
Crimson delila	13·15	12
White	93·51	80
Total*	374·00	374

theory, do not depend on the experimental data in any other way), there are 6 degrees of freedom ($n' = 7$); $\chi^2 = 9{\cdot}81$ and $P = 0{\cdot}13$; hence the deviations of experiment from expectation cannot be regarded as significant.

The Additive Nature of χ^2

4.3. A useful property of χ^2 is that several values can be added together, and if the degrees of freedom are also added, the total deviations of several distributions can be tested by entering the tables at the total degrees of freedom and finding the probability corresponding to the total χ^2. For instance, for the three distributions we have just tested, the total χ^2 is 21·30, the total degrees of freedom are 20, and P for this value of χ^2 and $n' = 21$ is 0·38; thus, taken alto-

* The figures given in the paper add up to 373·98, and we have added 0·01 of a unit to each of the largest groups to make the total correct. There are also two groups with expected frequencies of zero, but these must be left out in this test.

gether, the differences are still attributable to random errors. This is the correct way of combining the experiences of several distributions, and it is quite wrong to find an average P or χ^2.

To illustrate further this additive character of χ^2, Table 4.3 has been compiled from some data of Mendel, quoted by Bateson (1913), and gives the numbers of round and angular peas from ten plants,

TABLE 4.3

FREQUENCIES OF PEAS

Plant Number	Round n_s	Angular n_t	Total N	Ratio n_s/N	χ^2
1	45	12	57	0·789 5	0·47
2	27	8	35	0·771 4	0·09
3	24	7	31	0·774 2	0·10
4	19	10	29	0·655 2	1·39
5	32	11	43	0·744 2	0·00
6	26	6	32	0·812 5	0·67
7	88	24	112	0·785 7	0·76
8	22	10	32	0·687 5	0·67
9	28	6	34	0·823 5	0·98
10	25	7	32	0·781 2	0·17
Total ..	336	101	437	—	—
Expected	327·75	109·25	437·00	—	—

together with the ratios of the numbers of round peas to the totals. These ratios vary considerably, and we will see how far the variations may be explained on the hypothesis that the peas from the ten plants are effectively ten random samples from an infinite population in which the ratio is 0·75 (this is the Mendelian expectation). We might calculate for each plant the quantity

$$\frac{\textit{Deviation in Ratio from } 0{\cdot}75}{\textit{Standard Error}}$$

and since the sampling distribution of each ratio is binomial, this

$$\frac{\left(0{\cdot}75 - \frac{n_s}{N}\right)}{\sqrt{(1 - 0{\cdot}75)\frac{0{\cdot}75}{N}}}.$$

If the deviations were random, these ratios would be distributed approximately normally with unit standard deviation, and none of the ten would be expected to exceed 2·0. Alternatively we may calculate ten values of χ^2, and since there are two frequency groups, each plant contributes one degree of freedom. In this simple case, χ^2 happens to be the square of the above ratio.

Fisher gives the value of χ^2 lying on the 0·05 level of significance as 3·841 for one degree of freedom, and as none of those in Table 4.3 is as large as this, no individual plant differs significantly from expectation. Further, for all plants combined, the proportion of round peas is 0·768 9, giving a χ^2 of 0·96, which also is insignificant. It may be, however, that the variability in the proportion of round peas from plant to plant is greater than can be explained by random errors, when all are considered together. To make this clear, we may imagine an extreme case in which the values of χ^2 for all the plants are near the level of significance, but the ratio n_s/N varies above and below expectation, so that the ratio for all plants combined is near expectation. Then, although no *individual* plant appears to differ significantly from expectation, it is unlikely that all would be so near the level of significance if it were not that the deviations as a whole were real, but in different directions, so that when added they average out. To test such a point we may add the values of χ^2, giving a total of 5·30 for 10 degrees of freedom ($n' = 11$) so that $P = 0{\cdot}87$, and we conclude from the data that the plants do not vary significantly, and they may be regarded as so many random samples of peas. Had there been enough plants, instead of adding the values of χ^2 we could have formed a frequency distribution of them, and have compared it with the theoretical form for one degree of freedom.

The χ^2 test is thus an extension of the ordinary method of finding the significance of a single binomial mean, and enables the information given by a number of them to be combined.

4.31. It follows, of course, that if a number of values of χ^2 and their degrees of freedom may be added and treated as one, a total χ^2 may be split up into parts, and each part may be tested separately.

Contingency Tables

4.4. The problem of the sex-ratio of rats (in Table 3.5, p. 83) may be looked at in a different way. We are not concerned with the total numbers of males, of females, or of young resulting from the two

diets, but with the distribution of young in the four cells giving the two sex-ratios. If diet has had no effect on the sex-ratio, the 571 observations would be expected to be distributed at random in the four cells, with the one restriction that they should add up to give the totals of the table. In the infinite population of tables with those totals, the probability of an observation falling in the "deficient" group is 276/571, and that of it falling in the male group is 268/571, so that the probability of it falling in the "deficient" male square is

$$\frac{268}{571} \times \frac{276}{571},$$

and the expected number of individuals in that square is that probability multiplied by 571 = 129·54. Similarly, the other squares can

TABLE 4.4

EXPECTED FREQUENCIES

Diet	Males	Females	Total
Vitamin B deficient	129·54	146·46	276·00
Vitamin B sufficient.. ..	138·46	156·54	295·00
Total	268·00	303·00	571·00

be filled as in Table 4.4; they are the frequencies that would be expected if sex were independent of diet and the individuals were distributed at random. We may now test to see if Table 3.5 differs significantly from Table 4.4 by finding χ^2 for the four cells, and then the P for one degree of freedom. There is only one degree of freedom, since only one cell can be filled independently; the numbers in the others can be obtained from that one and the totals. In our example, $\chi^2 = 1{\cdot}20$ and $P = 0{\cdot}28$ (from Fisher's table); the deviations are not greater than can be attributed to random errors. It will be noticed that this probability is practically the same as that obtained previously from the standard error of the ratios (0·27); indeed, it should be, for both methods are equivalent, being based on the assumption that the number in a cell (or its ratio to the total) is distributed normally.

Table 3.5 is an example of a fourfold *contingency table*. When the individuals in a sample have two characters, and a frequency table is made classifying them according to both so as to show the relation

between the characters, the result is a contingency table. In Table 3.5, the two characters are sex and vitamin B content of diet, and complete information regarding these for any one rat is given by its position in the table. If the numbers in the cells are not randomly distributed, the two characters are said to be *associated*; the sex of the offspring in our example is not associated with the vitamin B content of the diet of the parent. Contingency tables may be manifold, and if there

TABLE 4.5

		Severity of Attack					
		Hæmorrhagic	Confluent	Abundant	Sparse	Very Sparse	Totals
Years since vaccination	0–10	— (0·87)	1 (5·13)	6 (9·02)	11 (8·60)	12 (6·38)	30
	10–25	5 (13·25)	37 (78·20)	114 (137·45)	165 (130·96)	136 (97·14)	457
	25–45	29 (27·04)	155 (159·47)	299 (280·32)	268 (267·07)	181 (198·10)	932
	over 45	11 (4·50)	35 (26·52)	48 (46·62)	33 (44·42)	28 (32·94)	155
Unvaccinated ..		4 (3·34)	61 (19·68)	41 (34·59)	7 (32·95)	2 (24·44)	115
Total ..		49	289	508	484	359	1 689

are n rows and m columns, there are $(n - 1)(m - 1)$ degrees of freedom.

Table 4.5 is a 5×5 contingency table, being Brownlee's data of the severity of smallpox attack and degree of vaccination (quoted by Pearson, 1910); below the frequencies, the expectations for independence are given in brackets.

As there are several expected frequencies less than ten in the first row and column, these have been combined with the second row and column in working out, so forming a 4×4 table. $\chi^2 = 196{\cdot}33$, there are 9 degrees of freedom, and P is less than $0{\cdot}000\,001$; hence the association between degree of vaccination and severity of attack is overwhelmingly significant. Notice, however, that this gives no information as to the strength of association; a smaller sample would

give a lower significance for the same association, and a larger sample would give a higher significance. Indeed, this test does not even tell us whether increased severity of attack is associated with a longer or a shorter period since vaccination. Having established, however, that the effect is real, we glance at the table again, and notice that for the more recently vaccinated patients there tends to be a deficit of severe attacks and an excess of milder ones; for the remotely vaccinated, the reverse is the case, and thus we establish the *sense* of the association; a measure of its *strength* will be discussed in section 9.5. Readers must be warned that association does not necessarily mean a causal relationship; it only means that in the population sampled there is a tendency for variations in the two characters to occur together. A much closer analysis is necessary to investigate causes; this also will be discussed later in section 7.3.

The test of Table 4.5 for association may be looked at in another way. The table consists effectively of a number of frequency distributions of years since vaccination, and χ^2 is the measure of the deviations of these distributions from hypothetical ones deduced from the "totals" column and differing only in total frequencies. The independent "constants" of these hypothetical distributions are four of the proportionate frequencies in the totals column; the fifth may be obtained from the others, since they all add up to unity. In addition to these, the five totals in the last row give the five totals of the hypothetical distributions, so that altogether nine "constants" have been fitted, leaving 16 degrees for a 5×5 table. Alternatively, the table may be regarded as a collection of five distributions in which the variate is the severity of attack.

For expressing and testing association, the contingency table is exceedingly useful, since the characters need only be described qualitatively, and need not be given in any particular order. A rearrangement of the rows or columns in Table 4.5 does not affect the result.

4.41. A useful application of the χ^2 test is to the comparison of a number of frequency distributions, e.g. for the purpose of checking sampling technique. The separate distributions may be regarded as rows in a contingency table, and if the χ^2 is large enough, they are significantly different. As a special case of this, when there are two

distributions and g' groups in each, there are $g' - 1$ degrees of freedom, and the expression for χ^2 becomes

$$\chi^2 = \underset{s}{S} \frac{N_1 N_2 \left(\frac{{}_1n_s}{N_1} - \frac{{}_2n_s}{N_2} \right)^2}{{}_1n_s + {}_2n_s} \quad . \quad . \quad . \quad . \quad (4.3)$$

where N_1 and N_2 are the two totals (they need not be equal),

${}_1n_s$ and ${}_2n_s$ are frequencies in one corresponding group in the two distributions,

and $\underset{s}{S}$ is the summation over all groups.

The grouping must, of course, be the same for both distributions under comparison. These tests are alternative to those involving only the means and standard deviations, but are more complete, since they compare the distributions in all respects.

General Notes on the Distribution of χ^2

4.5. The distribution of χ^2 is of considerable general importance in statistical theory, and a proper understanding of its nature is desirable. We have introduced it as it first appeared historically, as a measure of the differences between two frequency distributions. More generally, however, if we have any quantity x which is distributed normally about zero with unit standard deviation, and we add the squares of g independent values of x, the distribution of this sum in the infinite population is that of χ^2 for g degrees of freedom.

Thus, the distribution of χ^2 is closely related to that of N times the variance, s^2, in samples of N from a normal population in which the standard deviation is unity. Indeed, the sum of squares of g independent deviations from the population mean is equivalent to $g + 1$ deviations from the sample mean, so that if in equation (2.8), p. 66, we write

$$N = g + 1, \quad s^2 = \frac{\chi^2}{g + 1} \quad \text{and} \quad \sigma = 1$$

in the distribution of s, the equation reduces to the form of (4.2).

Tables of the probability integral of χ^2 only go up to $g = 30$, but for larger values of g it is sufficient to assume $\sqrt{2\chi^2}$ to be normally distributed about a mean of $\sqrt{2g - 1}$ with unit standard deviation

(Fisher, 1936*a*). Notice, however, that here we are asking not "is the observed $\sqrt{2\chi^2}$ significantly different from the mean $\sqrt{2g-1}$?" but rather "is the observed $\sqrt{2\chi^2}$ significantly greater than zero?" and consequently the deviation of $\sqrt{2\chi^2}$ which has a tail 5 per cent. of the *whole* curve is equivalent to the 5 per cent. level of significance. Such a deviation is 1·65 times the standard error, and $\sqrt{2\chi^2}-\sqrt{2g-1}$ has to be greater than 1·65 to be significant. When $g=30$, χ^2 lying on the 5 per cent. level on this assumption is 43·5, whereas Fisher's tables based on the correct distribution give 43·773, so the approximation is close enough for large samples.

CHAPTER V

SMALL SAMPLES

IT is not always possible to obtain large samples, such as are suitable for application of the methods of the last chapters, so the theory has been developed to give more exact methods suitable for small numbers. We stated in section 3.1 that there is nothing in the definition of a random sample that depends on its size, and would reiterate this in view of the existence of a fairly common impression to the contrary, of which the following quotation is typical.

"Moreover, in agricultural experiment, the number of observations rarely, if ever, is sufficiently large to allow full play to the laws of chance."

This is a misconception, for the laws of chance (as we know them) have full play every time a single individual is drawn from the population, and the theory of random sampling may be applied to small samples. There are two kinds of modification of the theory for large samples. The first consists in developments to the theory of estimation, which do not alter the general principles described in section 3.7. We need not concern ourselves with them any further. The second kind of modification is of practical importance and consists in using exact sampling distributions instead of the approximate ones used with large samples. We shall deal with this more fully.

Unfortunately, although the limitation of size has been removed that of form has not, and most of the following results apply only to quantities distributed normally.

VARIANCE ESTIMATED FROM SMALL SAMPLES

5.1. With a few observations, it is futile to form a frequency distribution, but the usual frequency constants may be calculated, and regarded as estimates, obtained from the sample, of the constants of the infinite population. For the normal population, the mean is found in exactly the same way as for large samples, but the best estimate of the variance (σ^2) is obtained by dividing the sum of the squares of the deviations from the mean, *not* by the number of observations, but by the number of *degrees of freedom*. Here the number of degrees of freedom is the number of deviations minus the number of constants determined from the sample and used to fix the points from

which those deviations are measured; in the simple case, when the mean only is found from the sample, the degrees of freedom are one less than the number of observations. We will illustrate this by the data of Table 1.3, which are the 100 random observations from an artificially constructed population, divided into groups of 5. We may find the variance either from the squares of the deviations from the grand mean or, regarding each group of five as a small sample, from the squares of the deviations from the sample means. For the latter process, instead of finding the variance for each sample separately, we may sum the squares of all the deviations and divide by the total degrees of freedom contributed by the ten samples. The sum of squared deviations from the grand mean is 8 864·75, and on dividing this by 99 we obtain the variance, viz. 89·5. The sum of squares from the sample means is 7 056·4, and since each sample contributes four degrees of freedom, there are 80 degrees altogether and the variance is 88·2. There is quite a fair agreement between the two estimates.* If we had used the old method for large samples we should have obtained values

$$\frac{8\,864\cdot75}{100} = 88\cdot6 \qquad \text{and} \qquad \frac{7\,056\cdot4}{100} = 70\cdot6,$$

with a much poorer agreement. In a large sample, the difference between dividing by N and $(N - 1)$ is quite unimportant.

As a special case, it may be determined that the mean variance for a number of pairs is

$$\frac{S(x_1 - x_2)^2}{2M}$$

where x_1 and x_2 are the individuals of any pair, M is the number of pairs and S is the summation over all pairs. In this way, the variance of any character between brothers from the same parents can be obtained as accurately from pairs of brothers from a hundred families as from a single family of one hundred and one brothers.

5.11. The justification for using the new estimate when the deviations are measured from the mean arises mathematically from the

* We have not tested the agreement by the ordinary sampling theory, using the standard errors of the values, since the values are not independent; they have been taken from the same data.

sampling distribution of the variance. If v' is the sample value of the variance as defined in section 1.23 and s is the corresponding standard deviation,

$$s = \sqrt{v'}, \quad ds = \frac{1}{2}\frac{dv'}{\sqrt{v'}}$$

and the sampling distribution of v' may easily be derived from that given for s in section 2.72; it is

$$df = K_2' v'^{\frac{N-3}{2}} e^{-\frac{Nv'}{2\sigma^2}} dv',$$

where K_2' is a constant and N is the size of sample. The mean value of v' obtained by integrating this distribution is

$$\bar{v}' = \frac{N-1}{N}\sigma^2.$$

Thus, as an estimate of σ^2, v' tends on the average to give a biassed value that is slightly smaller than the population value. It is preferable in dealing with small samples to use the unbiassed estimate,

$$v = \frac{Nv'}{N-1} = \frac{S(x-\bar{x})^2}{N-1}.$$

Significance of Means: The t Test

5.2. When testing the significance of the deviation of a sample mean from an assumed population value, we used the fact that the ratio of the deviation to its standard error is distributed normally with unit standard deviation.* If d is the deviation, σ is the population value of the standard deviation and N is the size of sample, this ratio is

$$\frac{d}{\frac{\sigma}{\sqrt{N}}}.$$

Now, in practice we do not always know σ and have to substitute the sample estimate, s. When dealing with large samples, s and σ are sufficiently alike to make this course reasonably accurate; but

* The ratio is the w of equation (2.5), section 2.51.

when the samples are small, some allowance must be made for the error involved in using s instead of σ.

The ratio by which we test the significance of a deviation may be written

$$t = \frac{d}{\frac{s}{\sqrt{N}}},$$

and if we know the sampling distribution of t, we may carry out the test exactly, without being limited to large samples. This was first pointed out by "Student" (1908) and he gave some tables of the sampling distribution of a quantity z closely related to t. These have now been generally superseded by tables of t given by "Student" (1925) and Fisher (1936*a*). When computing t for use with Fisher's tables, s must be taken as the square root of the variance calculated correctly from the sample by dividing the sum of squares of deviations by the degrees of freedom. Thus,

$$d = \frac{Sx}{N} - \bar{\xi} = \bar{x} - \bar{\xi}$$

$$s = \sqrt{\frac{S(x - \bar{x})^2}{N - 1}} \qquad \ldots\ldots (5.1)$$

where $\bar{\xi}$ is the population mean.

The sampling distribution of t is symmetrical and depends only on the number of degrees of freedom on which s has been estimated; it approaches the normal form as the degrees of freedom increase. Fisher's tables give, for different degrees of freedom, called n, values of t lying on various levels of significance, the levels being the sum of the two "tails" of the distribution as for the third column of our Table 2.5.

For large samples, $t = 2{\cdot}0$ lies on the 0·05 level of significance; other values that lie on the same level are: $t = 12{\cdot}7$ for samples of 2 ($n = 1$), $t = 4{\cdot}3$ for those of 3 ($n = 2$), $t = 2{\cdot}8$ for those of 4 ($n = 3$) and $t = 2{\cdot}3$ for those of 10 ($n = 9$); the effect of the non-normality of t is serious for samples much smaller than 20.

Table 5.1 shows the effect of a small electric current on the growth of maize seedlings, giving the difference between the elongation of the treated and untreated in parallel pairs of boxes (data from

Collins, Flint and McLane, 1929), a positive difference showing that the electrical treatment increased the rate of growth. The mean is 4·29 mm.; is this significantly different from zero? The sum of squares of deviations from the mean is 937·189, and since on our hypothesis $\bar{\xi} = 0$,

$$t = 4{\cdot}29\sqrt{\frac{90}{937{\cdot}189}} = 1{\cdot}33,$$

and $n = 9$; according to the tables, $t = 1{\cdot}38$ lies on the 0·2 level of significance, and there is thus no evidence from the sample that the treatment has made any difference to growth. The mean elongation is not large enough compared with the variations between those of the separate boxes to be significant.

TABLE 5.1

ELONGATION IN MM. (TREATED—UNTREATED).

6·0
1·3
10·2
23·9
3·1
6·8
− 1·5
− 14·7
− 3·3
11·1
Mean 4·29

Essential Character of t

5.21. Essentially, t is the ratio of a quantity distributed normally about a mean of zero, to an estimate based on n degrees of freedom, of the standard error of that quantity. Any such ratio, however it arises, has the same sampling distribution as t.

SIGNIFICANCE OF DIFFERENCES BETWEEN MEANS

5.3. The quantity t may be used for testing the significance of the difference between two sample means. We shall deal here only with the situation in which the assumption is made that the samples are from populations having a common standard deviation σ as well

as the same mean.* Then, if N_1 and N_2 are the numbers in the two samples, and $\bar{x}_1$ and $\bar{x}_2$ are the sample means, $(\bar{x}_1 - \bar{x}_2)$ is distributed normally about zero and an estimate of its standard deviation (or error) is

$$s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}},$$

where s is obtained by summing the squares of the deviations from the two sample means, dividing by the total degrees of freedom and finding the square root. Thus

$$s^2 = \frac{S_1(x_1 - \bar{x}_1)^2 + S_2(x_2 - \bar{x}_2)^2}{(N_1 - 1) + (N_2 - 1)} \quad . \quad . \quad . \quad (5.2)$$

where S_1 and S_2 are the summations over the two samples and x_1 and x_2 are individuals in the two samples. Then

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}, \quad . \quad . \quad . \quad . \quad . \quad (5.3)$$

which in large samples is distributed normally with unit standard deviation, is in small samples distributed as Fisher's t, the degrees of freedom being

$$n = (N_1 - 1) + (N_2 - 1).$$

Our example is from data provided by Corkill (1930) showing the effect of insulin on rabbits; the results for separate animals are given in Table 5.2. The difference in means is not large compared with the variations within each sample, and a statistical test of significance is necessary. The sums of squares of the deviations are 0·253 0 and 0·071 5, and thus

$$s^2 = \frac{0{\cdot}324\,5}{19} = 0{\cdot}017\,08 \quad \text{and} \quad s = 0{\cdot}130\,7;$$

* Since normality is also assumed, the hypothesis and assumptions amount to the composite hypothesis that the two populations are one. Judging from experience, however, the tests are not very sensitive to moderate departures from normality nor to small differences in standard deviation.

there are ten and eleven rabbits in the two samples, so

$$\sqrt{\frac{1}{N_1}+\frac{1}{N_2}} = 0{\cdot}436\,9$$

and

$$t = \frac{0{\cdot}138}{0{\cdot}130\,7 \times 0{\cdot}436\,9} = 2{\cdot}41.$$

TABLE 5.2

Muscle Glycogen (per cent.)	
Controls	After Insulin
0·19	0·15
0·18	0·13
0·21	Trace*
0·30	0·07
0·66	0·27
0·42	0·24
0·08	0·19
0·12	0·04
0·30	0·08
0·27	0·20
—	0·12
Means 0·273	0·135

* Assumed to be 0·00.

The mean variance has been calculated on 19 degrees of freedom, and for these conditions, $t = 2{\cdot}09$ lies on the 0·05 level of significance; thus the effect of insulin on muscle glycogen appears to be real.

The difference between this and the earlier example of the effect of electrical treatment on growth is that here there is no reason for taking controls and treated animals in pairs, they are *all* independent; in the former instance, boxes were treated in parallel pairs, and there was some reason for expecting that these would be subject to some of the same disturbing factors, which thus would not affect the differences.

SIGNIFICANCE OF DIFFERENCES BETWEEN VARIANCES

5.4. When testing the difference between variabilities for large samples in section 3.53, we assumed $(s_1 - s_2)$, the difference in standard deviations of the samples, to be distributed normally with a standard error of

$$\sqrt{\frac{\sigma^2}{2N_1} + \frac{\sigma^2}{2N_2}}$$

where σ is the standard deviation in the population from which the samples are presumed to be taken; and not knowing σ, we substituted s_1 and s_2 in the expression. Errors arising from this approximation are again important in small samples, but Fisher (1924*a* and 1936*a*) has suggested a new index:

$$z = \tfrac{1}{2}(\log_e s_1^2 - \log_e s_2^2) = \log_e \frac{s_1}{s_2}, \quad . \quad . \quad . \quad (5.4)$$

where s_1^2 and s_2^2 are the two variances calculated on the degrees of freedom. For samples of N_1 and N_2 drawn from the same population, if n_1 and n_2 are the numbers of degrees of freedom, z is distributed in the form

$$df = k\frac{e^{n_1 z}}{(n_1 e^{2z} + n_2)^{\frac{1}{2}(n_1+n_2)}}dz,$$

where n_1 and n_2 are the degrees of freedom $(= N_1 - 1$ and $N_2 - 1)$ and k is a constant.* This distribution, containing as variables only z, n_1 and n_2, is independent of the standard deviation of the population, σ, and since it involves no approximating assumptions, is applicable to small samples. The quantity z may vary between plus and minus infinity, being negative when s_1/s_2 is less and positive when s_1/s_2 is greater than unity, and unless $n_1 = n_2$, is skew. The positive part of the curve of $z = s_1/s_2$, however, is the same as the negative part of $z = s_2/s_1$, and so the probability integrals for positive deviations only are sufficient for any combination of degrees of freedom, the others can be obtained by interchanging n_1 and n_2. It is simpler, however, not to deal with negative values of z, but always to take the difference of logarithms so that it is positive, and hence to choose n_1 to be the degrees of freedom on which the *larger* variance

* This distribution of z is related to that of s and hence to that of χ^2 (see Fisher, 1924*a*).

is measured. Fisher (1936*a*) gives values of z at which ordinates cut off "tails" of 5, 1 and 0·1 per cent. of the total area of the curve for values of n_1 and n_2 chosen according to the above convention.* These points, however, are not the corresponding levels of significance as we understand them. The hypothesis is that the two variances are estimates of one and the same population value, and that z (which is the difference between the natural logarithms of the two standard deviations) is zero. Applying the arguments of section 3.32, we regard as lying on the 0·05 level of significance values of z at which ordinates cut off tails, each of which is about 0·025 of the whole area; the 5 per cent. level of significance, therefore, lies somewhere between the 5 and 1 per cent. points of Fisher's tables, unless there is some *a priori* reason for supposing the variance estimated by s_1^2 is either equal to or greater than that estimated by s_2^2 (compare p. 77).

When n_1 and n_2 are large, or when they are moderate and nearly equal, the distribution of z becomes nearly normal with a standard error of

$$\sqrt{\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

and in such circumstances a value of z greater than twice this value lies above the 5 per cent. level of significance. For example, when $n_1 = n_2 = 24$, this standard error becomes 0·204, and a z of 0·408 lies on the 5 per cent. level (that is between 0·342 5 and 0·489 0, at which according to Fisher's tables $P = 0{\cdot}05$ and 0·01). In order to check the assumption of normality, we may compare the deviation at which an ordinate cuts off a tail of 5 per cent. of the whole curve (1·65 times the standard error $= 1{\cdot}65 \times 0{\cdot}204 = 0{\cdot}337$) with 0·342 5; the agreement is quite good.

For the heights of Englishmen and Scotsmen of Table 3.1, p. 74, the standard deviations are 2·548 and 2·480, so $z = \log_e 1{\cdot}027 = 0{\cdot}026\,6$, and using the above approximation its standard error is

$$\sqrt{\frac{1}{2}\left(\frac{1}{6\,194}+\frac{1}{1\,304}\right)} = 0{\cdot}021\,5;$$

* Mahalanobis (1932) gives similar tables for the 5 and 1 per cent. points of s_1/s_2 and s_1^2/s_2^2, and by using these, the rather troublesome transformation of equation (5.4) may be avoided. Tables D1 and D2 at the end of this book give corresponding values of $\log_{10} s_1^2/s_2^2$.

z is only 1·23 times its standard error and is not significant. The other test in section 3.53 gave the difference in standard deviations as 1·27 times its standard error, and this shows that in very large samples it is as well to use the difference between standard deviations as z.

As an illustration of a small sample, we will test the difference in variability of muscle glycogen between the controls and insulin-treated rabbits of Table 5.2. The sums of squares are 0·253 0 and 0·071 5, the variances are 0·028 11 and 0·007 15, $z = \frac{1}{2}\log_e 3{\cdot}93 = 0{\cdot}684$, and $n_1 = 9$ and $n_2 = 10$.

Fisher's tables do not give z for these values of n_1 and n_2, but interpolation is made easy by the fact that for any one value of n_2, changes in z are nearly proportional to $1/n_1$. Linear interpolation with respect to $1/n_1$ is therefore appropriate. We shall write z_5 and z_1 for values on the 5 and 1 per cent. points.

Then, for $n_2 = 10$,

when $n_1 = 8$,

$$\frac{1}{n_1} = 0{\cdot}125\,00, \qquad z_5 = 0{\cdot}561\,1, \qquad \text{and} \quad z_1 = 0{\cdot}810\,4;$$

when $n_1 = 12$,

$$\frac{1}{n_1} = 0{\cdot}083\,33, \qquad z_5 = 0{\cdot}534\,6, \qquad \text{and} \quad z_1 = 0{\cdot}774\,4;$$

and the differences are:

$$diff.\left(\frac{1}{n_1}\right) = 0{\cdot}041\,67, \quad diff.(z_5) = 0{\cdot}026\,5, \quad \text{and} \quad diff.(z_1) = 0{\cdot}036\,0.$$

Hence, when $n_1 = 9$,

$$\frac{1}{n_1} = 0{\cdot}111\,11, \quad z_5 = 0{\cdot}561\,1 - \frac{0{\cdot}013\,89}{0{\cdot}041\,67} \times 0{\cdot}026\,5 = 0{\cdot}552\,3,$$

$$\text{and} \quad z_1 = 0{\cdot}810\,4 - \frac{0{\cdot}013\,89}{0{\cdot}041\,67} \times 0{\cdot}036\,0 = 0{\cdot}798\,4.$$

The value $z = 0{\cdot}684$ lies between the 5 and 1 per cent. points, but it is doubtful if it is on the 5-per cent. level of significance, and we can only say that although the data suggest that the effect of insulin has been to make the muscle glycogen percentages more regular, further observations are necessary to establish the fact.

Interpolation with respect to n_2 is scarcely ever necessary, since the tables are given for n_2 increasing by units between 1 and 30.

Significance of Variations between Several Samples

5.5. The method of section 5.3 tests the hypothesis that two samples are from populations having the same mean, assuming they have the same standard deviation. When there are more than two samples, the hypothesis may best be tested by the analysis of variance described in the next chapter.

It is sometimes useful to test the significance of the general differences between the variances of more than two samples. For instance, it may be desired to combine data from several experiments or sources, and before doing so, it is advisable to ascertain that the variance within each set of data is sensibly the same; or again, when examining the products of a factory statistically, any difference in the variability of the articles produced on different machines, sections, or times is taken as evidence of lack of control.

Neyman and E. S. Pearson (1931) have proposed an index for testing this. If there are k estimates of variance, $v_1\, v_2 \ldots v_k$, each based on N observations ($n = N - 1$ are the degrees of freedom), the index is*

$$L_1 = \frac{(v_1\, v_2 \ldots v_k)^{\frac{1}{k}}}{\frac{1}{k}(v_1 + v_2 + \ldots v_k)} \quad \ldots \quad (5{\cdot}5)$$

Thus, L_1 is the ratio of the geometric to the arithmetic mean of the variances. Mahalanobis (1933) and Nayer (1936) have tabled values of L_1 lying on the 5 and 1 per cent. levels of significance for various values of k and N.† Nayer's tables are fuller than the others. One property of L_1 is that it decreases in value as the variances differ more among themselves; it has a maximum value of unity when the variances are equal.

In a series of six experiments (Tippett, 1934), the following six variances of cotton yarn breakages were obtained: 0·2056, 0·5578, 0·6489, 0·3378, 0·6194 and 0·4311; each was based on 9 degrees

* Our notation differs from that of Neyman and Pearson.

† Mahalanobis and Nayer write n for the number of observations in each sample, corresponding to our N. In applying this test, we shall consistently use the small letter for degrees of freedom. Nayer writes f for our n.

of freedom. Before combining the results of these experiments, it was considered necessary to test the variances for homogeneity. Here, $k = 6$, $N = 10$ and $L_1 = 0{\cdot}930$. From Nayer's tables, $L_1 = 0{\cdot}81$ lies on the 5 per cent. level, and remembering that a high L_1 corresponds to greater uniformity of variance, we conclude that the value of $0{\cdot}93$ is not significant.

Small Samples from the Binomial and Poisson Distributions

5.6. In Chapter II we used the binomial and Poisson series to test the homogeneity of a large number of counts of seeds and yeast cells; but it is not always possible to obtain such large samples from one population. However, if there are small samples of replicate counts from a variety of populations, the experience so obtained can be reduced to a common measure and added.

5.61. We will develop the tests for the binomial by reference to Table 4.3. The numbers of round and angular peas and the totals may be regarded as forming a 2×10 contingency table, and obtaining the expectations from the totals (i.e. ignoring the Mendelian expectation), it may be tested for randomness by means of the χ^2 calculated on 9 degrees of freedom as shown in section 4.4. If this value of χ^2 is significant, the data are not random nor homogeneous. If there are several such tables obtained (say) from a number of varieties of peas in which the expectations are not necessarily equal, all the values of χ^2 and their degrees of freedom may be added, and the sum tested in the usual way. Further, there may be enough tables with the same number of degrees of freedom (say a hundred or more) to form a frequency distribution of χ^2, and this may be compared with the theoretical form obtained from the standard tables. The constant χ^2 is thus the common measure to which a variety of experiences may be reduced.

The expression for χ^2 becomes particularly simple if the number of peas from each plant is constant. Using more general language, if there are g' parallel sets, each of n trials, with m_1, m_2 . . . successes in the sets and a mean of $\bar{m}$ successes,* then

$$\chi^2 = \frac{S(m_s - \bar{m})^2}{\left(\bar{m} - \frac{\bar{m}^2}{n}\right)} \quad . \quad . \quad . \quad . \quad . \quad (5.6)$$

* In Table 4.3 the g' sets are the 10 plants, the n trials are the total peas on each plant (we are now dealing with the case where n is constant), and the m_1, m_2 . . . successes are the numbers of round peas.

and there are $(g' - 1)$ degrees of freedom. This, like the χ^2 obtained in section 4.3 under a slightly different assumption, is a measure of the variation in m_s between plants. If we had a number of such sets of 10 for different kinds of plants, we could find the χ^2 for each set, and compare their distribution with the theoretical one for 9 degrees of freedom. In the same way, for instance, the uniformity of conditions in seed germination can be investigated from a large experience of many sorts of seed, as long as there is the same number of replicates (g') on each seed.

In one series of experiments made to find the effect of various treatments on cotton seeds, 100 seeds were exposed in two lots of 50 for each treatment. In such a case, if m_1 and m_2 are the numbers germinating in the two lots of 50,

$$\chi^2 = \frac{(m_1 - m_2)^2}{(m_1 + m_2)\left(1 - \frac{m_1 + m_2}{100}\right)}$$

and there is one degree of freedom $(g' = 2)$. There were 52 treatments, giving a total χ^2 of $82 \cdot 84$. The tables do not extend to so many degrees of freedom, so we may use the approximation mentioned in section 4.5, assuming

$$\sqrt{2\chi^2} - \sqrt{2g - 1}$$

to be normally distributed with unit standard deviation. Here $g = 52$, and

$$\sqrt{2\chi^2} - \sqrt{2g - 1} = 2 \cdot 72,$$

corresponding to $P = 0 \cdot 003$; this suggests that there must have been some lack of uniformity in conditions. The frequency distribution of χ^2 and the theoretical form as found from Fisher's table for one degree of freedom are compared in Table 5.3; as Fisher gives only a few probability integrals, the intervals of χ^2 are unequal. There are obviously too many pairs of counts with large values of χ^2, and to test the divergences we must use the rather overworked constant, χ^2, again; but as shown in section 4.1, calculating a χ'^2 according to equation (4.1) as a measure of the differences between the two distributions of Table 5.3.* In doing this, we have com-

* This double use of χ^2 may be a little confusing, but if readers can for the moment forget that the variate of Table 5.3 is χ^2, and just regard the table as two distributions, the differences between which are to be tested by calculating χ'^2, all will be well.

bined the groups in pairs according to the brackets, and obtain $\chi'^2 = 7{\cdot}86$, which for three degrees of freedom (i.e. four combined groups and only the total of the theoretical distribution has been adjusted) lies almost exactly on the 0·05 level of significance. This test of randomness is less stringent than the one above adding all the values for χ^2, for it groups all those over 2·706 together, and takes no account of how much greater they may be.

TABLE 5.3

χ^2	Frequencies	
	Expected	Observed
0–0·015 8	5·2	5
0·015 8–0·064 2	5·2	1
0·064 2–0·148	5·2	3
0·148 –0·455	10·4	10
0·455 –1·074	10·4	10
1·074 –1·642	5·2	5
1·642 –2·706	5·2	8
over 2·706	5·2	10
Total	52·0	52

5.62. For the Poisson distribution, the index corresponding to that of equation (5.6) is

$$\chi^2 = \frac{S(m_s - \bar{m})^2}{\bar{m}}$$

where m_s is any individual count, $\bar{m}$ is the mean, and S is the summation over all counts. Fisher, Thornton and Mackenzie (1922) first used this to test the accuracy of bacterial counts and found significant deviations from expectation. More recently, Smith and Prentice (1929) have used it to check their technique of cyst counts in soil. They had 73 sets of 10 counts from different samples of soil, and calculated the 73 values of χ^2. The distribution of this is compared with the theoretical one for 9 degrees of freedom below in Table 5.4, and the expectations, being obtained from Fisher's book (1936*a*), are for unequal intervals of χ^2.

TABLE 5.4

χ^2	Frequencies	
	Expected (ν_s)	Observed (n_s)
0– 4·168	7·3	7
4·168– 5·380	7·3	7
5·380– 6·393	7·3	6
6·393– 8·343	14·6	22
8·343–10·656	14·6	18
10·656–12·242	7·3	2
12·242–14·684	7·3	5
over 14·684	7·3	6
Total	73·0	73

The agreement is quite fair; it is tested in the usual manner by calculating another

$$\chi'^2 = \sum_s \frac{(n_s - \nu_s)^2}{\nu_s} = 9{\cdot}604,$$

which for seven degrees of freedom gives P greater than 0·2.

CHAPTER VI

THE ANALYSIS OF VARIANCE

PREVIOUS chapters are based on the presentation of collections of data in the form of frequency distributions and on the description of a few features, particularly the extent and form of the variation, by frequency constants. This is a process of summarisation in which some detail is inevitably lost. We shall now reverse the process and deal with statistical methods designed to recover some of the detail that is of value, starting from the frequency distribution and its constants as bases. Comparatively little has been done in this direction with the *form* of variation, but the *amount* of variation can be split up into parts associated with different causes or sources. The method of analysis of variation is introduced in its simple quantitative form in this chapter; and most of the subjects dealt with in subsequent chapters are developed from the same basic method. As a first step, we shall in the next section prove a fundamental mathematical property of variance.

VARIANCE AN ADDITIVE QUANTITY

6.1. Variability may be considered to arise from a multitude of causes producing small deviations; but it often happens that to these there are added larger deviations due to a few more important causes, and the variation is not always homogeneous. It is a fundamental property of that measure of the degree of variability, the variance, that it is additive, i.e. if a quantity is subject to the operation of several independent causes each of which contributes a certain variance, then the final variance of the quantity is the sum of those due to the several causes. Let it be assumed, for example, that a quantity x is subject to random variations, and to others associated with two factors A and B; then the value of any one observation of x is

$$x = \bar{\xi} + \alpha + \beta + \xi',$$

where $\bar{\xi}$ is the mean, α and β are the deviations arising from A and B, and ξ' is the random deviation. The square of the deviation of x from its mean is

$$(x - \bar{\xi})^2 = \alpha^2 + \beta^2 + \xi'^2 + 2\alpha\beta + 2\alpha\xi' + 2\beta\xi',$$

and this may be summed for a sample of N individuals, and divided by the degrees of freedom (N in this case, since we have not found the mean $\bar{\xi}$ from the sample, but have assumed it). Thus we obtain

$$\frac{S(x-\bar{\xi})^2}{N} = \frac{S\alpha^2}{N} + \frac{S\beta^2}{N} + \frac{S\xi'^2}{N} + \frac{2S\alpha\beta}{N} + \frac{2S\alpha\xi'}{N} + \frac{2S\beta\xi'}{N}$$

and as N becomes indefinitely large, the last three terms of this equation tend to zero if α, β and ξ' are independent; the other terms are the squares of the standard deviations or variances, so that finally

$$\sigma_x^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\xi'}^2 \quad . \quad . \quad . \quad . \quad . \quad (6.1)$$

Hence the variance of x is the sum of the random variance and of those due to A and B. We used this property in Chapter III to find the standard error of a difference in terms of those of the two means; since the variance is the square of the standard error, the above equation leads directly to equation (3.1).* We shall use it now to analyse the variability of quantities into parts.

Analysis of Variance

6.2. Consider Table 6.1 in which the data (Harris, 1910) are frequency distributions of ovaries containing different numbers of ovules, and are for ten separate shrubs of the American Bladder Nut. The separate columns are called *arrays*. It is obvious that there are considerable differences between the shrubs, for while shrub 11 has ovaries with 22–30 ovules, shrub 13 has ovaries with between 17 and 24 ovules, and the other shrubs show similar differences. The variations between ovaries on any one shrub are less than those between all taken together, and the whole table suggests that we may legitimately divide the total variability into two parts: one associated with differences between ovaries from the same shrub, and another with differences between shrubs. We say that there is an association between the shrubs and ovules per ovary, and as evidence adduce the fact that the mean variance of deviations from the shrub means as found by the method of section 5.1, which is 3·057, is very much less than the variance of the "totals" column (5·385).

Indeed, Table 6.1 is really a manifold contingency table, but the association is expressed differently because one variate is quantitative;

* α and β may be positive or negative.

a treatment by the method of the contingency table would be very laborious with so many squares.

We may present the analysis of the variability into the two parts in a systematic manner. Let x be any individual reading, $\bar{x}_1\ \bar{x}_2 \ldots \bar{x}_s \ldots \bar{x}_m$ the m shrub means, $\bar{x}$ the grand mean, and n the number of readings per shrub, so that there are $nm = N$ readings altogether.

TABLE 6.1

FREQUENCIES OF OVARIES (SERIES O, 1908 C).

Ovules per Ovary	Serial Number of Shrub: 11	12	13	15	16	18	19	20	21	22	Totals
17	—	—	1	—	—	—	—	—	—	—	1
18	—	2	23	—	1	6	1	—	—	1	34
19	—	5	13	—	1	4	5	5	—	1	34
20	—	8	27	—	1	4	1	5	2	2	50
21	—	10	4	1	1	5	2	7	1	1	32
22	4	21	18	—	7	13	9	12	6	2	92
23	5	17	7	3	9	16	15	25	8	15	120
24	41	35	7	16	42	48	44	39	61	68	401
25	21	2	—	25	14	3	18	6	10	8	107
26	9	—	—	21	9	1	4	1	4	2	51
27	9	—	—	8	5	—	1	—	3	—	26
28	4	—	—	8	3	—	—	—	2	—	17
29	3	—	—	12	3	—	—	—	2	—	20
30	4	—	—	5	4	—	—	—	1	—	14
31	—	—	—	1	—	—	—	—	—	—	1
Totals ..	100	100	100	100	100	100	100	100	100	100	1 000

Then, for any ovary on any one shrub, the deviation from the grand mean is the deviation from the shrub mean plus that of the shrub mean from the grand mean:

$$(x - \bar{x}) = (x - \bar{x}_s) + (\bar{x}_s - \bar{x}).$$

We will square these deviations, and sum for all observations from the one shrub; denoting this summation by S' and applying the rules in section 1.311 we obtain

$$S'(x - \bar{x})^2 = S'(x - \bar{x}_s)^2 + n(\bar{x}_s - \bar{x})^2 + 2(\bar{x}_s - \bar{x})S'(x - \bar{x}_s).$$

The term $(x_s - \bar{x})^2$ is constant for the one shrub and its sum for the n observations is n times the single value (the second term in the above equation); the third term above is zero, since the sum of the deviations from the mean, $S'(x - \bar{x}_s)$, is necessarily zero. To obtain the total sum of squares, we now sum the terms of this equation again for all shrubs (using the sign $\underset{s}{S}$) and finally obtain

$$\underset{s}{S}S'(x - \bar{x})^2 = \underset{s}{S}S'(x - \bar{x}_s)^2 + n\underset{s}{S}(\bar{x}_s - \bar{x})^2 \quad . \quad . \quad . \quad (6.2)$$

The sign $\underset{s}{S}S'$ is equivalent to S, the simple summation over all observations in the sample. For the data of Table 6.1, equation (6.2) is

$$5\,379{\cdot}775 = 3\,026{\cdot}350 + 2\,353{\cdot}425,$$

and each of these terms is given an appropriate place in Table 6.2 of the Analysis of Variance. The terms are the sums of squares of deviations, (*a*) of individual observations from the grand mean, (*b*) of individual observations from the shrub means, and (*c*) of the shrub means from the grand mean (the sum being multiplied by n in this case). The degrees of freedom are given in the table; for the total, one mean has been found and there are thus 999 degrees; for the deviations from the shrub means, each shrub contributes 99 degrees, giving a total of 990; and for the shrub means, 10 deviations are measured from the grand mean, leaving 9 degrees of freedom. The sums of squares and degrees of freedom of the parts should add up to give the "total," and the sums of squares, divided by the degrees of freedom, give estimates of the variances. That for the "total" is the ordinary variance of all the observations in the sample, and that for "within a shrub" is the variance of the intra-shrub deviations found after the method of section 5.1, and is an estimate, based on 990 degrees of freedom, of the true variance, σ_r^2 (say). The variance "between shrubs" is n times the square of the standard deviation of shrub means. To investigate this more fully, let us suppose we have an indefinitely large number of ovaries from each of an indefinitely large number of shrubs, and let the variance of these shrub means be σ_s^2; this is the true value for the infinite population. If now we have only n ovaries from each shrub, the means are subject to random errors due to the intra-shrub variation, and their standard error is $\sigma_r/\sqrt{n}$. These errors increase the variability of the

shrub means, and the resulting variance may be obtained by adding the two components, as shown in equation (6.1), giving

$$\left(\sigma_s^2+\frac{\sigma_r^2}{n}\right);$$

the variance of Table 6.2 is an estimate based on 9 degrees of freedom of n times this. If v_s is the variance between shrubs, and v_r the variance within a shrub, as found from the sample

$$\left.\begin{aligned} v_s &\rightarrow n\sigma_s^2+\sigma_r^2 \\ v_r &\rightarrow \sigma_r^{2*} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (6.3)$$

These two relations are the basis of the expression of association. If the variation between shrubs is relatively important, σ_s^2 is large

TABLE 6.2

ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between shrubs ..	2 353·425	9	261·492
Within a shrub ..	3 026·350	990	3·057
Total	5 379·775	999	5·385

compared with σ_r^2, and the two estimates v_s and v_r are very different. If the shrub variation is zero, $\sigma_s^2=0$ and v_s tends to equal v_r. In such a case, for normally distributed variates, v_s and v_r are two independent estimates of the same variance, σ_r^2. They are subject to the same random errors as are all estimates of variance, and the significance of any difference should be tested by the methods of section 5.4. For Table 6.2, the one (261·492) is so much greater than the other (3·057), that the reality of the association needs no test. Using the relations of equation (6.3)

$$261\cdot492 \rightarrow 100\,\sigma_s^2+\sigma_r^2,$$

$$3\cdot057 \rightarrow \sigma_r^2,$$

whence

$$2\cdot584 \rightarrow \sigma_s^2.$$

* The sign $\rightarrow$ denotes that the quantity on the left is an estimate of that on the right, and that the former approaches the latter as the size of the sample (number of degrees of freedom in both parts) increases indefinitely.

If a sample of ovaries were taken at random from these shrubs, the number from each shrub being left to chance, the variance would be $\sigma_s^2 + \sigma_r^2$, of which an estimate is 5·641. This differs slightly from the variance in the "total" row of Table 6.2 because the sample to which the latter refers is not entirely random; it has been arranged so that 100 ovaries come from each shrub.

Table 6.3 summarises the above relations and sets out the analysis of variance for the general case of n observations on each of m arrays. The variance within an array is often called the residual; having

TABLE 6.3

ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between arrays	$n\underset{s}{S}(\bar{x}_s - \bar{x})^2$	$m - 1$	$v_s \to n\sigma_s^2 + \sigma_r^2$
Within an array	$\underset{s}{SS'}(x - \bar{x}_s)^2$	$m(n-1) = N - m$	$v_r \to \sigma_r^2$
Total ..	$S(x - \bar{x})^2$	$mn - 1 = N - 1$	—

performed the analysis, we are not very interested in the variance of the "total" row, since the two parts contain all the useful information.

COMPUTATION

6.3. When computing the sums of squares for Table 6.3, the various deviations may be found explicitly, and squared and added. If the arithmetic is correct, the sums of squares between and within arrays should add up to give the total; this provides a check. Usually, however, this process is laborious and it is convenient to apply a modification of equation (1.2), p. 37. If the original units are used so that $h = 1$, the appropriate part of equation (1.2) may be written

$$S(x - \bar{x})^2 = Sx^2 - N\bar{x}^2 = Sx^2 - \frac{T^2}{N}, \quad . \quad . \quad (6.4)$$

where S, $\bar{x}$ and N have their usual meanings and T is the total value of the variate for the N observations, i.e. $T = Sx = N\bar{x}$.

Applying this to equation (6.2) and writing $T = \underset{s}{S}S'x$ and $T_s = S'x$ for the sth shrub,* we have

$$\underset{s}{S}S'(x - \bar{x})^2 = \underset{s}{S}S'x^2 - \frac{T^2}{N},$$

$$\underset{s}{S}S'(x - \bar{x}_s)^2 = \underset{s}{S}S'x^2 - \frac{\underset{s}{S}T_s^{\,2}}{n}, \quad . \quad . \quad . \quad (6.5)$$

and

$$n\underset{s}{S}(\bar{x}_s - \bar{x})^2 = \frac{\underset{s}{S}T_s^{\,2}}{n} - \frac{T^2}{N}.$$

The process of computation may be written in words:

(1) Square the individual values and add,

(2) Find the total value of the variate for the arrays, square, add, and divide the result by the number of individuals per array, and

(3) Find the total value of the variate for all individuals, square, and divide by the grand total number of individuals.

Then if (1), (2) and (3) represent the results of the above operations, equation (6.2) becomes

$$[(1) - (3)] = [(1) - (2)] + [(2) - (3)].$$

When the data are grouped, readers should have no difficulty in applying equation (1.3), p. 38. In such instances, it is better not to use Sheppard's corrections, but to keep the grouping fairly fine so that they are unimportant. The precise effect of these corrections is not certain; but they increase the apparent association by reducing the residual variance, so to neglect them in testing significances is to be on the safe side.

It may be convenient to measure the values of the variate as deviations from some arbitrary origin and to divide them by an arbitrary constant h before summing and squaring. Then after applying equations (6.5) it is only necessary to multiply the resultant sums of squares by h^2. The process of computation based on equations (6.5) may easily be performed exactly, with the aid of a table of squares, down to the last stages of dividing by the numbers of

* These values of T are not the total numbers of individuals but are the total amounts of variate in the several arrays. In Table 6.1, T_s is the total number of ovules in the 100 ovaries from the sth shrub.

individuals, and since there are only two such divisions, these may be performed with ample accuracy without much labour.

As an example, we have analysed the variance of the 100 readings of Table 1.3, p. 32, into two portions, between groups of 5 and within a group. The results are in Table 6.4. Since the groups are random samples $\sigma_s^2 = 0$, and the two estimates of variance should be nearly equal, as indeed they are; actually the variance within a group is less than that between groups, and we will test the significance of this difference. In the notation of section 5.4,

TABLE 6.4

ANALYSIS OF VARIANCE (DATA TABLE 1.3)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between groups ..	1 808·35	19	95·18
Within a group ..	7 056·40	80	88·20
Total	8 864·75	99	—

$z = \frac{1}{2}[\log_e 95{\cdot}18 - \log_e 88{\cdot}20] = 0{\cdot}038\,1$, $n_1 = 19$ and $n_2 = 80$.

From Fisher's table we see that when $n_1 = 24$ and $n_2 =$ infinity, a z of 0·208 5 lies on the 0·05 point, so our z of 0·038 1 is quite insignificant, and the two estimates of variance are equivalent.

TABLES WITH NON-UNIFORM ARRAY TOTALS

6.4. The analysis of variance may be performed on tables in which the numbers of individuals in the arrays are not equal. If the numbers in the arrays are $n_1\ n_2 \ldots n_s \ldots n_m$, equation (6.2) becomes

$$\underset{s}{S}S'(x - \bar{x})^2 = \underset{s}{S}S'(x - \bar{x}_s)^2 + \underset{s}{S}n_s(\bar{x}_s - \bar{x})^2 \quad . \quad (6.6)$$

and $\underset{s}{S}(T_s^2/n_s)$ is written for $\underset{s}{S}T_s^2/n$ in equations (6.5).

In such instances, the relations of equations (6.3) do not hold, for in summing the squares of the deviations of the group means from the grand mean, each group has been given a different weight, n_s. If the true variance between groups (σ_s^2) is zero, however, $v_s \longrightarrow v_r$, and the test for the existence of association is that v_s and v_r are significantly different.

TABLE 6.5

Length of Cuckoos' Eggs in mm. (central values)	Nest Type						Totals
	Meadow Pipit	Tree Pipit	Hedge Sparrow	Robin	Pied Wagtail	Wren	
19·65	1	—	—	—	—	—	1
19·85	—	—	—	—	—	1	1
20·05	1	—	—	—	—	1	2
20·25	—	—	—	—	—	1	1
20·45	—	—	—	—	—	—	—
20·65	1	—	—	—	—	—	1
20·85	1	—	1	—	—	3	5
21·05	—	1	—	1	1	3	6
21·25	—	—	—	—	—	1	1
21·45	—	—	—	—	—	1	1
21·65	3	—	1	—	—	—	4
21·85	3	1	—	1	3	—	8
22·05	10	1	1	3	1	3	19
22·25	8	—	—	1	—	1	10
22·45	3	1	—	2	1	—	7
22·65	2	1	—	1	1	—	5
22·85	4	—	1	—	—	—	5
23·05	1	—	4	5	2	—	12
23·25	2	3	—	1	1	—	7
23·45	1	2	1	—	1	—	5
23·65	1	1	—	—	—	—	2
23·85	1	1	3	1	—	—	6
24·05	—	3	1	—	3	—	7
24·25	1	—	—	—	—	—	1
24·45	1	—	—	—	—	—	1
24·65	—	—	—	—	—	—	—
24·85	—	—	—	—	1	—	1
25·05	—	—	1	—	—	—	1
Totals n_s	45	15	14	16	15	15	120
T_s ..	56	78	75	42	64	— 69	246
$\frac{T_s^2}{n_s}$ ..	69·69	405·60	401·78	110·25	273·07	317·40	504·30

Table 6.5 gives distributions of the lengths of cuckoos' eggs found in the nests of a variety of other birds (Latter, 1902). The total frequencies for the kinds of nest vary, and since the observations are fewer, the association is not quite so obvious as in the previous example. The grouping is rather fine for such a small sample, but has been adopted because Sheppard's corrections will not be applied. The values have been measured as deviations from the length 22·05 mm. in terms of the unit of grouping 0·2 mm. The quantities T_s and T_s^2/n_s are given in these units at the foot of the table, and we may calculate that

$$\underset{s}{S}S'x^2 = 3\,934, \quad \underset{s}{S}\frac{T_s^2}{n_s} = 1\,577{\cdot}79 \quad \text{and} \quad \frac{T^2}{N} = 504{\cdot}30.$$

TABLE 6.6

ANALYSIS OF VARIANCE (LENGTHS OF CUCKOOS' EGGS)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between nest types ..	1 073·49	5	214·7
Within a nest type ..	2 356·21	114	20·7
Total	3 429·70	119	—

From these, the terms of equation (6.6) are found and inserted in Table 6.6. The variance between nest types is greater than the residual, and we will test it for significance. We find

$$z = \tfrac{1}{2}[\log_e 214{\cdot}7 - \log_e 20{\cdot}7] = 1{\cdot}17, \quad n_1 = 5 \quad \text{and} \quad n_2 = 114.$$

For $n_1 = 5$ and $n_2 = 60$, $z = 0{\cdot}779\,8$ lies on the 0·1 per cent. point, so we must conclude that the association between egg-length and type of nest is real. These variances are in terms of the arbitrary units (0·2 mm.), and if they are required in mm.^{2} they must be multiplied by 0·04.

RELATION BETWEEN z AND t TESTS

6.5. A special case of analysis occurs when there are two groups; this is the comparison between two samples. If there are n_1' and n_2'

observations in each group, $\bar{x}_1$ and $\bar{x}_2$ are the means, and $\bar{x}$ is the grand mean; the analysis of variance is as set out in Table 6.7.

S_1 and S_2 are the summations over samples 1 and 2, and S is the summation over the whole; also, if v_s is greater than v_r,

$$z = \tfrac{1}{2}[\log_e v_s - \log_e v_r] = \log_e \sqrt{\frac{v_s}{v_r}}$$

and the degrees of freedom are $n_1 = 1$ and $n_2 = n_1' + n_2' - 2$.

The square root of the ratio of variances is the t of equation (5.3), p. 115, which was used in testing the difference between two means. Thus, if we make the transformation

$$z = \tfrac{1}{2}\log_e t^2 \quad \text{or} \quad t = e^z,$$

TABLE 6.7

ANALYSIS OF VARIANCE (TWO GROUPS)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between samples	$n_1'(\bar{x}_1-\bar{x})^2+n_2'(\bar{x}_2-\bar{x})^2 = \dfrac{(\bar{x}_1-\bar{x}_2)^2}{\left(\frac{1}{n_1'}+\frac{1}{n_2'}\right)}$	1	v_s
Within a sample	$S_1(x-\bar{x}_1)^2+S_2(x-\bar{x}_2)^2$	$n_1' + n_2' - 2$	v_r
Total ..	$S(x-\bar{x}^2)$	$n_1' + n_2' - 1$	—

the distributions of t for n degrees of freedom, say, and of z for $n_1 = 1$ and $n_2 = n$ are equivalent. This is because z and t are mathematically related; for any value of t there is only one value of z.

A slight modification in interpretation of the z test is necessary before we can show its equivalence to that based on t. For the simple use of z to test the difference between two variances, we showed in section 5.4 that the 0·05 level of significance is about the 2·5 per cent. point; on the other hand, when used as an alternative to t, the 0·05 point of z is also the 0·05 level. When testing two variances, the hypothesis is that in the population $z = 0$, and a positive or negative deviation, if sufficiently large, may be incompatible with this,

showing that the variances are really different.* Now when t is zero, the corresponding z is minus infinity, and when t is plus infinity, z is also plus infinity; hence the whole distribution of z is equivalent to the positive half only of the distribution of t. Applying the argument of section 3.32, if t is always made positive, the 0·05 level of significance is the deviation of t at which an ordinate cuts off a tail 0·05 of the area of the positive half of the curve of t; consequently, if we use the curve of z instead, the deviation beyond which the tail is 0·05 of the whole curve is the 0·05 level of significance (since the whole of the z curve is equivalent to the positive half of that for t). The difference between the two treatments of z lies in the type of question asked. In the former use we ask if z is significantly different from zero, and a large negative value may be so; in the latter use for comparing two means we ask if z is significantly greater than minus infinity (i.e. if t is greater than zero), and no negative value can be so. A large negative value of z is quite compatible with the hypothesis that the means of the two samples are really equal, for it only shows that t is smaller than would be expected, and that the samples are a little too much alike. These varying distinctions between 0·05 points and levels of significance are very puzzling, but they should be clearly understood.

We find from the tables that when $n = 24$, $t = 2{\cdot}064$ lies on the 0·05 level, whence $z = \log_e 2{\cdot}064 = 0{\cdot}725$ for $n_1 = 1$ and $n_2 = 24$ should also lie there. Fisher's tables of z give the value 0·724 6 as the 0·05 point.

Significances of Groups of Means

6.6. This process of analysis of variance is the better method of testing the differences between several means as a whole, referred to in section 5.5, and is eminently suited to small samples; for the distributions of z which are used are exact, provided the form of the residual deviations is normal and their variance is constant for all means. When Fisher's tables of z are used to test the significance of differences between several means, the 0·05 point is also the same level of significance, as we have just shown for pairs of means.

* The fact that we interchange the degrees of freedom, choosing n_1 to be associated with the larger variance and keeping z positive, is purely a matter of computing convenience, and avoids the necessity of having duplicate tables with negative values of z. The complete distribution of z contains negative deviations from zero.

Table 6.8 is taken from a paper by Warren (1909), and shows the mean head breadths of numbers of termites (small soldiers) taken from five nests during five months.*

We will first analyse the variance into two parts, between nests and within a nest. The actual deviations from the grand and nest means have been squared and summed and the sums are entered in Table 6.9. The between-nest variance is much greater than its residual, giving a z of 0·829, which lies beyond the 0·01 point ($z = 0{\cdot}744\,3$ lies on it); so the variation from nest to nest is real. This merely confirms what would be expected from an examination of Table 6.8, for nests 670 and 675 have the highest means in all

TABLE 6.8

MEAN HEAD BREADTHS OF TERMITES (SMALL SOLDIERS) IN MM.

Nest Number ..	668	670	672	674	675	Means
November ..	2·273	2·479	2·404	2·447	2·456	2·411 8
January ..	2·332	2·603	2·457	2·388	2·626	2·481 2
March ..	2·375	2·613	2·452	2·515	2·633	2·517 6
May ..	2·373	2·557	2·396	2·445	2·487	2·451 6
August ..	2·318	2·377	2·279	2·312	2·410	2·339 2
Means ..	2·334 2	2·525 8	2·397 6	2·421 4	2·522 4	2·440 3

months, 672 and 674 nearly always come next, and 668 has the lowest mean in all but the month of August.

The reality of the differences between months is not quite so clear, for although August has the lowest mean in all nests but one, and March the highest in all but another, the other months show no consistent differences. We can, however, test this in the same way by analysing the total variance into two parts, as in the second part of Table 6.9. The "between months" variance is certainly greater than the residual, but

$$z = \tfrac{1}{2}\log_e\left(\frac{0{\cdot}023\,51}{0{\cdot}008\,72}\right) = 0{\cdot}496,$$

* This sort of table must not be confused with one like Table 6.1. Here the entries are measurements, there they are frequencies; here the characters are three in number (head breadth, nest number, and month of year), there they are only two (ovules per ovary and shrub number).

and being barely on the 0·05 point ($z = 0·526\,5$ is just on it) is of doubtful significance; hence on the simple analysis, the seasonal variation, though suggestive, is not well established. The situation is complicated, however, by the nest variation, and this treatment is not quite sufficient; more complete methods will be dealt with in section 10.3.

6.7. We have shown that the method of the analysis of variance well expresses the heterogeneity of variability. This heterogeneity is obvious when the effect is great, but the analytical method is

TABLE 6.9

ANALYSIS OF VARIANCE (HEAD BREADTHS OF TERMITES)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Between nests	0·137 442	4	0·034 36
Within a nest	0·130 967	20	0·006 55
Between months ..	0·094 046	4	0·023 51
Within a month ..	0·174 363	20	0·008 72
Total	0·268 409	24	—

objective, and so is particularly useful when the effect is smaller and not so obvious. Like all quantitative methods, statistical or otherwise, it merely specifies objective measures for qualities which may be subjectively appreciated in extreme cases. So far, however, we have not been concerned with more than testing for the existence of association, and have paid no attention to the problem of expressing its strength. As in all such tests, the probability of significance, depending as it does on the size of the sample, gives no indication of that strength.

Although the methods and tests described in this chapter are only applicable when the distribution within each array is normal, work by E. S. Pearson and others, published in *Biometrika*, indicates that moderate skewness, such as is apparent in Table 6.1, is not likely to lead to serious error.

It is further assumed that the variance within the separate arrays is constant within the limits of the errors of random sampling. The residual variance is a mean variance for all arrays. Where non-uniformity is suspected, its significance may be tested by the method of section 5.5.

CHAPTER VII

CORRELATION

Correlation Tables and Scatter Diagrams

7.1. When dividing the total variability of a quantity into parts, one of which is associated with arrays as in Table 6.1, it is by no means inevitable that the character which defines the arrays should be some qualitative description or serial number; it also may be the groups of some quantitative variate. Tables 7.1 to 7.5 are examples in which the individuals are classified according to two quantitative variates; these are called *correlation tables*. In the first of these, the arrays are groups of eggs having a small sub-range of length; each array is regrouped according to longitudinal girth. The association between the two quantities is well marked, and it is clear that there is also a tendency for the length to increase with the girth quite regularly. We have thus reached another stage in the treatment of variability; for while the single distribution shows the extent and form of the variation and the table of arrays introduced in the last chapter shows the association of parts of the variation with other factors, now the correlation table discovers the nature of that association when the other factor is a quantitative variate. It will be noticed, however, that such a table may be approached in two ways; either variate may be regarded as the one which is being analysed, and both the columns and rows are arrays.

When arranging correlation tables, it is convenient to choose the grouping so that there are from ten to twenty for each variate; if any observation appears to fall exactly on the dividing line between two groups, a half may be given to each, and similarly it may happen that a quarter of a unit may be assigned to each of four adjacent cells in the table. In specifying the characters, either the sub-ranges of the groups may be given, as in Tables 7.1 and 7.4, or the central values as in Table 7.5. The actual process of making a correlation table may be carried out in two ways. The first is to mark the position of each observation in the table by means of a dot or a stroke and finally to count the marks. This, however, is only suitable for fairly small samples, for if one makes a mistake or loses the place, there is nothing for it but to start again; further, the only means of checking the accuracy of the table is to repeat it. The

second method is to write the values of the characters of each individual on a separate card, and to sort the cards; the first sorting is into groups according to one character, and then each group is resorted according to the second character. It is then an easy matter to look through each pile of cards to see that none is out of place, and finally to count them and enter the numbers in a table. The writing of the cards takes some time, but it is a saving in the long run,

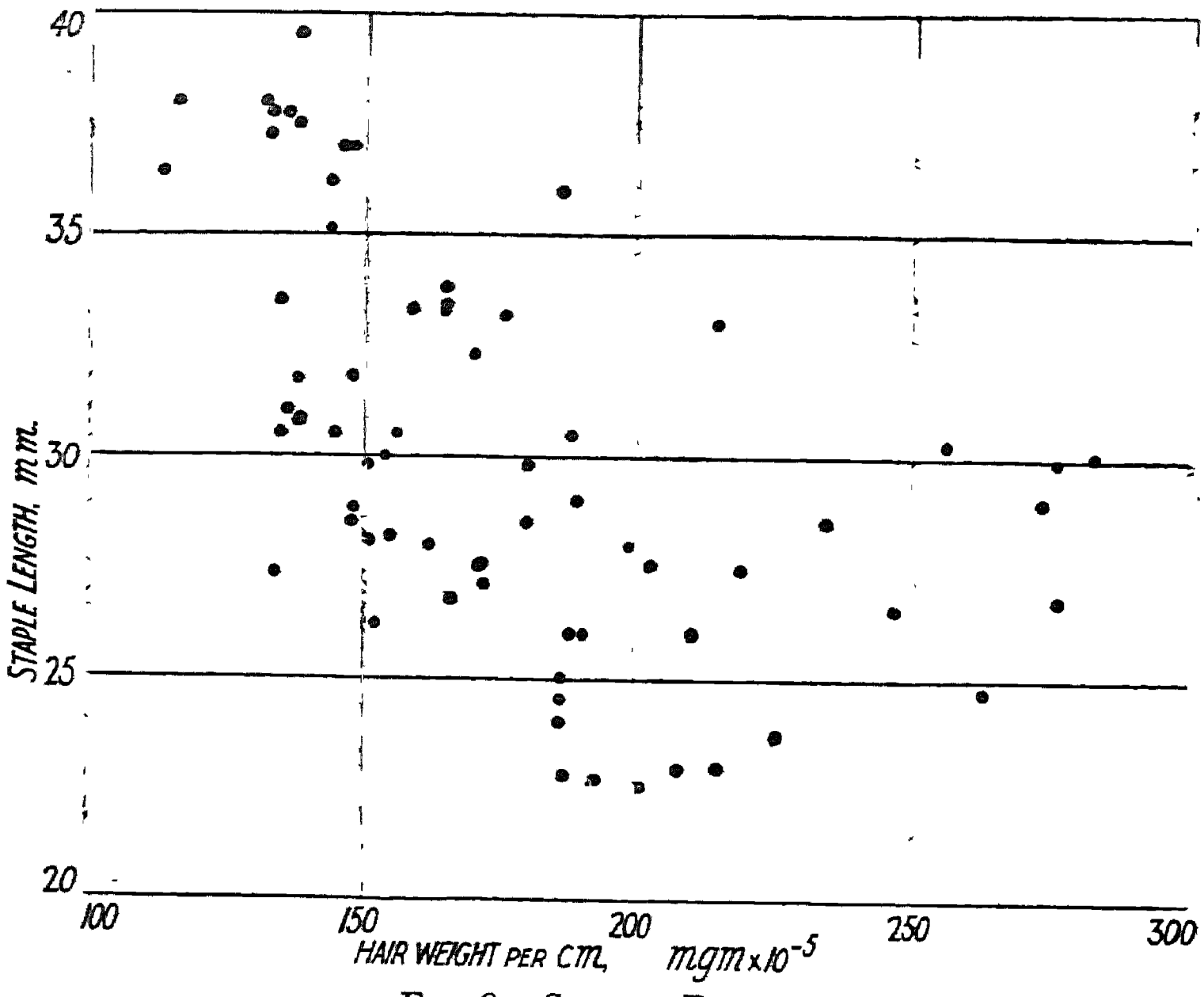


FIG. 8.—SCATTER DIAGRAM.

particularly if there are more than two characters; it is often possible to collect the data on cards and so avoid copying.

If the observations are few, a correlation table is not suitable, and its place may be taken by a *scatter diagram*. This is just an ordinary graph in which x and y are the two variates, and points on this represent observations; Fig. 8 is an example, and shows the relation between length and fineness of the hairs of a number of varieties of cotton.* If the relationship between the two variates is exact, the points in the scatter diagram lie on a smooth curve, while as the relationship weakens, the diagram more and more

* Data by Morton (1926).

TABLE 7.1

Length and Longitudinal Girth of Eggs of the Common Tern

$r = +0{\cdot}89$

(Data from "A Co-operative Study," 1923)

Longitudinal Girth in cm.		Length in cm.										
		3·55–	3·60–	3·65–	3·70–	3·75–	3·80–	3·85–	3·90–	3·95–	4·00–	4·05–
		–11	–10	–9	–8	–7	–6	–5	–4	–3	–2	–1
9·80–	–14	1	—	—	—	—	—	—	—	—	—	—
9·90–	–13	—	—	—	—	—	—	—	—	—	—	—
10·00–	–12	—	—	—	—	—	—	—	—	—	—	—
10·10–	–11	—	—	—	—	—	—	—	—	—	—	—
10·20–	–10	—	—	—	—	—	—	—	—	—	—	—
10·30–	– 9	—	—	—	1	—	—	—	1	—	—	—
10·40–	– 8	—	—	—	—	1	—	1	—	—	—	—
10·50–	– 7	—	—	—	—	2	2	1	—	—	—	1
10·60–	– 6	—	—	—	—	1	6	2	4	2	—	—
10·70–	– 5	—	—	—	—	—	2	4	—	3	1	—
10·80–	– 4	—	—	—	—	1	3	3	8	9	8	2
10·90–	– 3	—	—	—	—	1	1	1	11	7	11	5
11·00–	– 2	—	—	—	1	—	—	1	2	20	20	23
11·10–	– 1	—	—	—	—	—	—	1	3	6	20	20
11·20–	0	—	—	—	—	—	—	1	1	7	6	21
11·30–	1	—	—	—	—	—	—	—	—	—	—	7
11·40–	2	—	—	—	—	—	—	—	—	—	2	3
11·50–	3	—	—	—	—	—	—	—	—	—	—	1
11·60–	4	—	—	—	—	—	—	—	—	1	—	—
11·70–	5	—	—	—	—	—	—	—	—	—	—	2
11·80–	6	—	—	—	—	—	—	1	—	—	—	—
11·90–	7	—	—	—	—	—	—	—	—	—	1	—
12·00–	8	—	—	—	—	—	—	—	—	—	—	—
12·10–	9	—	—	—	—	—	—	—	—	—	—	—
12·20–	10	—	—	—	—	—	—	—	—	—	—	—
12·30–	11	—	—	—	—	—	—	—	—	—	—	—
Totals	..	1	—	—	2	6	14	16	30	55	69	85

TABLE 7.1—*continued*

LENGTH AND LONGITUDINAL GIRTH OF EGGS OF THE COMMON TERN

$r = + 0 \cdot 89$

(*Data from "A Co-operative Study,"* 1923)

Length in cm.												
4·10–	4·15–	4·20–	4·25–	4·30–	4·35–	4·40–	4·45–	4·50–	4·55–	4·60–	4·65–	Totals
0	1	2	3	4	5	6	7	8	9	10	11	
—	—	—	—	—	—	—	—	—	—	—	—	1
—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	2
—	—	—	—	—	—	—	—	—	—	—	—	2
—	—	—	—	—	—	—	—	—	—	—	—	6
—	—	—	—	—	—	—	—	—	—	—	—	15
—	—	—	—	—	—	—	—	—	—	—	—	10
1	—	—	—	—	—	—	—	—	—	—	—	35
1	—	—	—	—	—	—	—	—	—	—	—	38
10	1	1	—	—	—	—	—	—	—	—	—	79
25	8	3	1	2	1	—	—	—	—	—	—	90
37	29	11	5	—	—	1	—	—	—	—	—	119
38	25	23	13	5	1	1	1	—	—	—	—	114
10	25	29	21	9	2	—	—	—	—	—	—	101
5	11	20	27	22	9	1	2	—	—	—	—	98
1	2	8	15	30	12	5	1	1	—	—	—	76
—	—	2	9	25	16	10	2	—	—	—	—	66
—	—	—	—	9	10	15	11	1	1	—	—	48
—	—	1	—	4	2	5	4	6	6	—	—	29
—	—	—	—	—	1	2	4	5	3	1	—	16
—	—	—	—	—	—	—	—	1	1	1	3	6
—	—	—	—	—	—	—	—	1	1	—	1	3
—	—	—	—	—	—	—	—	—	—	—	1	1
128	101	98	91	106	54	40	25	15	12	2	5	955

resembles the result that would be obtained by sprinkling the points evenly with a pepper-pot; the former type of diagram is usually the experience of the physicist. The scatter diagram and correlation table are equivalent, except that in the latter the area is divided into a number of rectangles, and the number of points in each one counted and substituted by the numerals. It is well to choose the scales of a diagram so that the range of variation is about the same in both directions, and the diagram is nearly square.

The Correlation Coefficient

7.2. Correlation tables may show all degrees of association from the obvious one between the length and girth of eggs in Table 7.1 to the inappreciable one between head length and reaction time to sight in Table 7.5. This quality is measured by the *correlation coefficient* (r), and the corresponding value is given with each of Tables 7.1 to 7.5; it is small when the association is weak, and large when the relationship is fairly close and strong. We shall now proceed to discuss the meaning of this coefficient and related constants, postponing a description of the methods of estimation and computation to a later part of the chapter. The correlation coefficient may be regarded from three points of view, discussed in the following sections headed Frequency Surfaces, Regression Lines and Analysis of Variance.

Frequency Surfaces

7.21. Just as a single frequency distribution may be represented graphically by a histogram, so a correlation table may be represented by a similar figure in three dimensions. The base is divided into a number of rectangles representing the cells of the table, by a number of lines parallel to the x- and y-axes, and on these rectangles are raised columns, proportional in volume to the frequency in the corresponding cell of the table. Such is an empirical *frequency surface*, and it must be noted that frequencies are measured by volumes.

Such a surface, of course, is made up of step-like figures (rather reminiscent of the basaltic columns of the Giant's Causeway in Northern Ireland) and shows irregularities; but as the size of the sample is increased the irregularities disappear, and as the number of groups is increased the steps become smaller, until in the limit a smooth surface (analogous to the frequency curve for a single variate)

TABLE 7.2

Numbers of Pistils and Stamens in *Ranunculus Ficaria*

Late Flowers. $r = +0{\cdot}75$

(Data by Weldon, 1901)

Number of Pistils	Number of Stamens: 8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	Totals
5	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
6	—	—	1	1	—	1	—	3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6
7	—	—	—	2	4	4	2	4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	16
8	—	—	1	—	—	3	10	14	2	1	3	1	—	—	—	—	—	—	—	—	—	—	35
9	—	—	2	—	1	2	3	7	10	2	7	—	1	—	—	—	—	—	—	—	—	—	35
10	—	—	—	—	—	1	3	4	6	5	13	4	1	—	—	—	1	—	—	—	—	—	38
11	1	—	—	—	—	—	2	5	8	6	6	7	5	—	—	—	—	—	—	—	—	—	40
12	—	—	—	1	—	—	2	4	4	3	5	7	6	3	—	—	—	—	—	—	—	—	35
13	—	—	—	1	—	—	1	—	4	7	13	9	6	2	—	1	—	—	—	1	—	—	45
14	—	—	—	—	—	—	—	1	—	2	8	7	12	4	2	—	—	—	—	—	—	—	36
15	—	—	—	—	—	—	—	1	1	2	5	4	2	1	3	1	1	—	—	—	—	—	21
16	—	—	—	—	—	—	—	—	—	2	3	—	5	3	5	1	1	2	1	—	—	—	23
17	—	—	—	—	—	—	—	—	—	1	3	3	2	4	1	—	—	2	—	—	—	—	16
18	—	—	—	—	—	—	—	—	—	—	—	—	1	2	2	2	—	—	3	—	—	1	11
19	—	—	—	—	—	—	—	—	—	—	—	1	1	—	2	1	1	1	1	—	1	—	9
20	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1	—	—	2
21	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	1
22	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	1
23	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	1
24	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1
Totals	1	—	4	6	5	11	23	43	35	31	66	43	42	19	17	7	4	6	6	2	1	1	373

TABLE 7.3.—NUMBERS OF PISTILS AND STAMENS IN *Ranunculus Ficaria Early Flowers.* $r = +0.51$

Number of Pistils	Number of Stamens: 18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	Totals
2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
9	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	2
10	—	—	—	2	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	3
11	—	—	3	1	2	1	1	3	1	1	—	—	—	—	—	—	—	—	—	—	—	13
12	—	1	1	1	4	1	1	1	—	—	1	1	—	—	—	—	—	—	—	—	—	12
13	—	4	3	—	1	2	4	1	2	3	1	1	—	—	—	—	—	—	—	—	—	22
14	—	—	1	2	4	3	7	4	4	5	2	1	—	1	1	—	—	—	—	—	—	35
15	1	—	—	2	—	—	1	5	3	5	4	5	3	2	—	—	—	—	—	—	—	31
16	—	—	—	—	2	2	1	4	2	3	1	2	5	2	1	—	—	—	—	—	—	25
17	—	—	—	1	—	2	4	3	1	7	1	3	2	1	—	2	—	—	—	—	—	27
18	—	—	—	—	—	1	2	—	5	4	4	1	1	1	1	1	—	—	—	—	—	21
19	—	—	—	—	2	—	1	1	2	2	—	1	3	4	2	1	—	—	—	—	—	19
20	—	—	—	—	1	—	—	—	2	2	—	—	2	4	1	1	—	—	—	—	—	13
21	—	—	—	—	—	—	—	1	2	3	—	1	2	3	1	2	—	—	—	—	—	15
22	—	—	—	—	—	—	—	—	1	1	—	2	1	1	2	—	—	2	—	—	—	10
23	—	—	—	—	—	—	—	—	1	1	—	2	—	—	—	—	—	—	—	—	—	4
24	—	—	—	—	—	—	—	—	—	—	—	—	—	1	2	—	—	—	—	1	—	4
25	—	—	—	—	—	—	—	—	—	—	—	—	1	—	2	—	—	—	—	—	—	3
26	—	—	—	—	—	—	—	1	—	1	—	—	—	—	—	—	—	2	—	—	—	4
27	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1
28	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
29	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
31	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Totals	1	6	8	9	16	12	22	26	26	38	14	23	20	20	13	7	1	4	—	1	1	268

TABLE 7.4

Daily Maximum and Minimum Temperatures at Rothamsted for August 1878–1926

(Data by Fisher and Hoblyn, 1928.)* $r = +0{\cdot}30$

Maximum Temperature ° F.	Minimum Temperature ° F. 36–	38–	40–	42–	44–	46–	48–	50–	52–	54–	56–	58–	60–	62–	Totals
50–	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
52–	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
54–	—	—	—	—	—	1	3	2	—	—	—	—	—	—	6
56–	—	—	1	3	4	2	4	3	3	1	—	—	—	—	21
58–	1	1	—	3	7	10	11	11	4	2	1	—	—	—	51
60–	—	1	4	2	11	13	24	22	13	8	4	—	—	—	102
62–	—	2	3	10	9	24	29	37	35	16	8	3	—	—	176
64–	—	2	3	6	16	25	43	39	54	28	12	3	—	—	231
66–	—	1	3	10	9	16	28	38	40	37	18	7	2	—	209
68–	2	—	2	9	16	19	20	33	35	35	33	12	2	—	218
70–	1	2	1	6	9	9	10	15	14	32	20	13	4	1	137
72–	—	—	1	3	6	10	9	14	19	19	25	14	5	—	125
74–	—	—	3	3	4	7	12	6	13	11	7	10	3	2	81
76–	—	—	—	—	5	3	8	4	7	8	8	7	3	1	54
78–	—	—	—	2	2	3	5	7	9	9	3	2	3	1	46
80–	—	—	—	—	1	2	—	3	5	8	3	5	1	—	28
82–	—	—	—	—	—	—	2	1	4	1	3	2	—	—	13
84–	—	—	—	1	—	1	—	—	1	—	2	2	1	—	8
86–	—	—	—	—	—	—	—	—	1	2	1	—	1	—	5
88–	—	—	—	—	—	—	—	1	—	—	1	1	—	—	3
90–	—	—	—	—	—	—	—	—	—	—	1	—	1	—	2
92–	—	—	—	—	—	—	—	—	—	—	1	—	—	—	1
Totals ..	4	9	21	58	100	145	208	236	257	217	151	81	26	5	1 518

* This table is condensed from that of the original paper by doubling the sub-ranges of the arrays.

TABLE 7.5—Reaction Time to Sight and Head Length

(Data by Harmon, 1926.) Corrected for Age. $r = -0{\cdot}03$

Reaction Time to Sight in 1/100 Sec.	Head Length in Inches																			Totals
	6·805	6·905	7·005	7·105	7·205	7·305	7·405	7·505	7·605	7·705	7·805	7·905	8·005	8·105	8·205	8·305	8·405	8·505	8·605	
5·995	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1
7·995	—	—	—	—	—	1	—	1	1	—	—	—	—	—	—	—	—	—	—	3
9·995	—	—	1	2	1	5	5	10	10	12	12	7	6	4	1	—	—	—	—	76
11·995	—	—	—	1	1	1	5	9	10	15	15	5	8	6	4	2	1	—	—	83
13·995	—	—	2	4	6	10	23	29	30	49	39	27	19	9	7	1	—	—	—	255
15·995	—	—	4	2	14	37	67	115	131	138	145	121	71	39	17	6	1	2	—	910
17·995	—	5	4	11	22	43	76	143	191	201	177	146	83	52	22	11	1	2	—	1 190
19·995	2	1	5	13	37	58	122	148	176	196	215	138	81	63	31	9	3	—	2	1 300
21·995	—	1	1	7	11	25	41	53	76	56	78	49	28	25	8	5	—	1	—	465
23·995	—	—	1	—	7	10	24	22	58	45	42	24	19	11	5	1	—	1	—	270
25·995	—	—	—	3	3	5	13	13	13	12	14	5	5	6	2	—	—	—	—	94
27·995	—	—	1	—	1	—	1	5	5	2	4	2	2	2	—	—	—	—	—	25
29·995	—	—	—	—	1	—	—	—	1	4	2	1	2	1	—	—	—	—	—	12
31·995	—	—	—	—	—	—	—	1	—	—	1	—	—	—	—	—	—	—	—	2
33·995	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	1
35·995	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	1
37·995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
39·995	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	1
41·995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
43·995	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1
Totals ..	2	7	19	43	104	196	378	550	702	731	745	525	324	218	97	35	6	6	2	4 690

results. The progress in devising systems of mathematical formulæ to describe such surfaces is less than has been made for the single variate distributions, and methods based on the *normal surface* are applied to most distributions. The equation to the normal surface is

$$df = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{(1-r^2)}}\, e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}\right)} dxdy,$$

where x and y are the two variates measured as deviations from their means, σ_x and σ_y are constants equal to the standard deviations of the two variates, r is a constant equal to the correlation coefficient, N is the total number in the sample, and df is the element of frequency in the range $dxdy$ with its centre at (x, y). If we cut this surface by any vertical plane parallel to either the x- or y-axis, the section is a normal frequency curve of standard deviation $\sigma_x\sqrt{1-r^2}$. or $\sigma_y\sqrt{1-r^2}$, Fig. 9 shows the contours of surfaces in which correlation coefficients are 0, 0·3, 0·6 and 0·9, and the standard deviations of x and y are both equal to unity; it will be seen that the surfaces all rise to a hump at the centre, but that they tend also to form a diagonal ridge which becomes narrower and sharper as r increases, as would be expected from the fact that the standard deviation of a sectional curve becomes smaller as r approaches unity. The correlation coefficient can never be greater than unity, and as it approaches that value, the ridge tends to become a thin outline in the form of a normal curve running diagonally; when $r = 0$, vertical planes through the centre parallel to the x- and y-axes divide the surface into four equal quadrants. The correlation coefficient can be negative, but the only difference that makes is to cause the major axes of the ellipses to follow the other diagonal.

Regression Lines

7.22. An easier way of treating a correlation table is to draw a curve relating the means of the arrays of x and y. This can be done in two ways; either we may find the mean of y for every array of x (i.e. referring to Table 7.1, find the mean girths when the lengths of the eggs are successively 3·55–3·60, 3·70–3·75, 3·75–3·80, etc., cm.) or we may find the mean of x for every array of y (mean lengths when the girths are successively 9·80–9·90, 10·30–10·40, 10·40–10·50, etc., cm.). In Fig. 10 are given the array means for each of

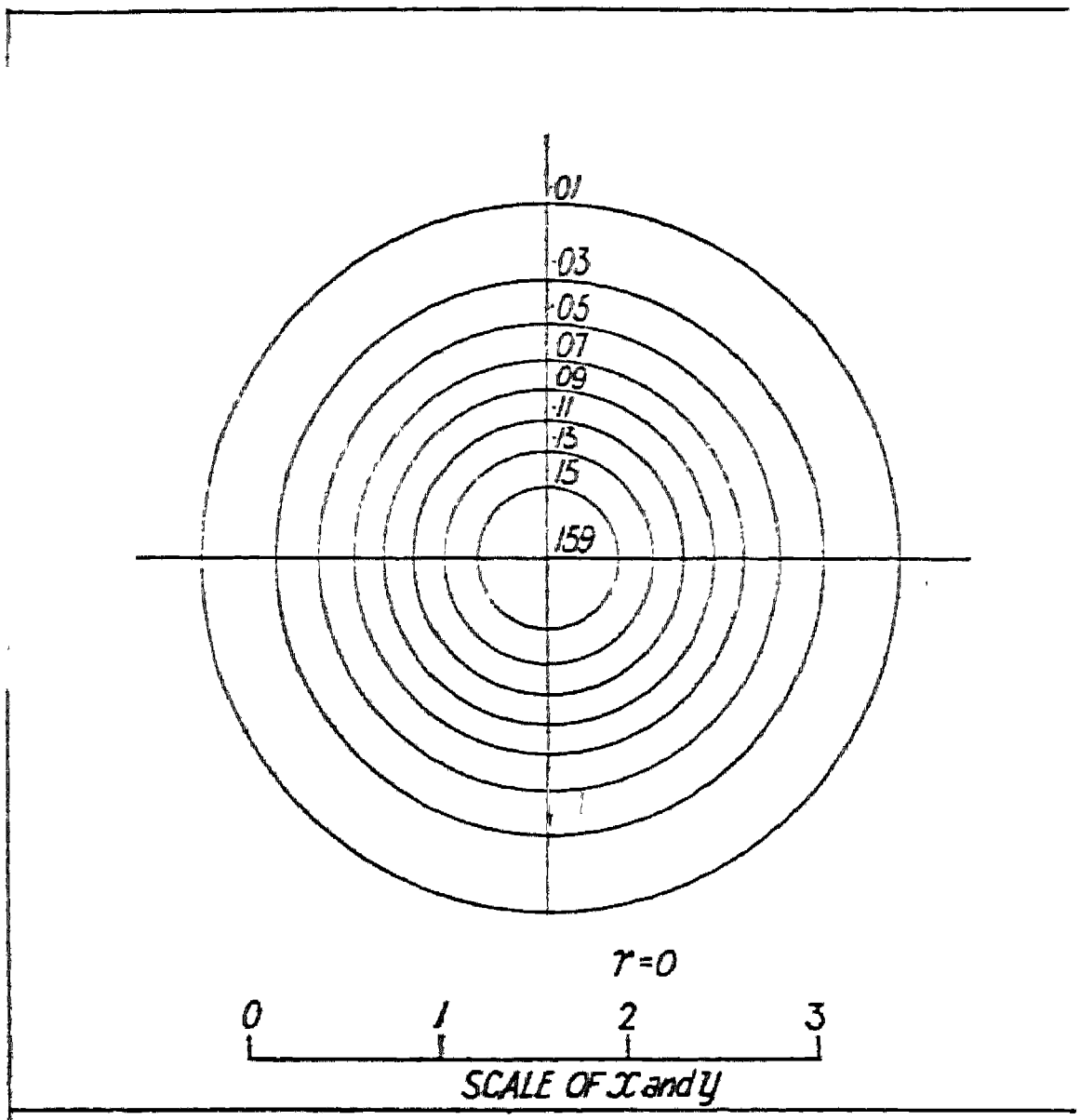


FIG. 9.—NORMAL FREQUENCY SURFACES.

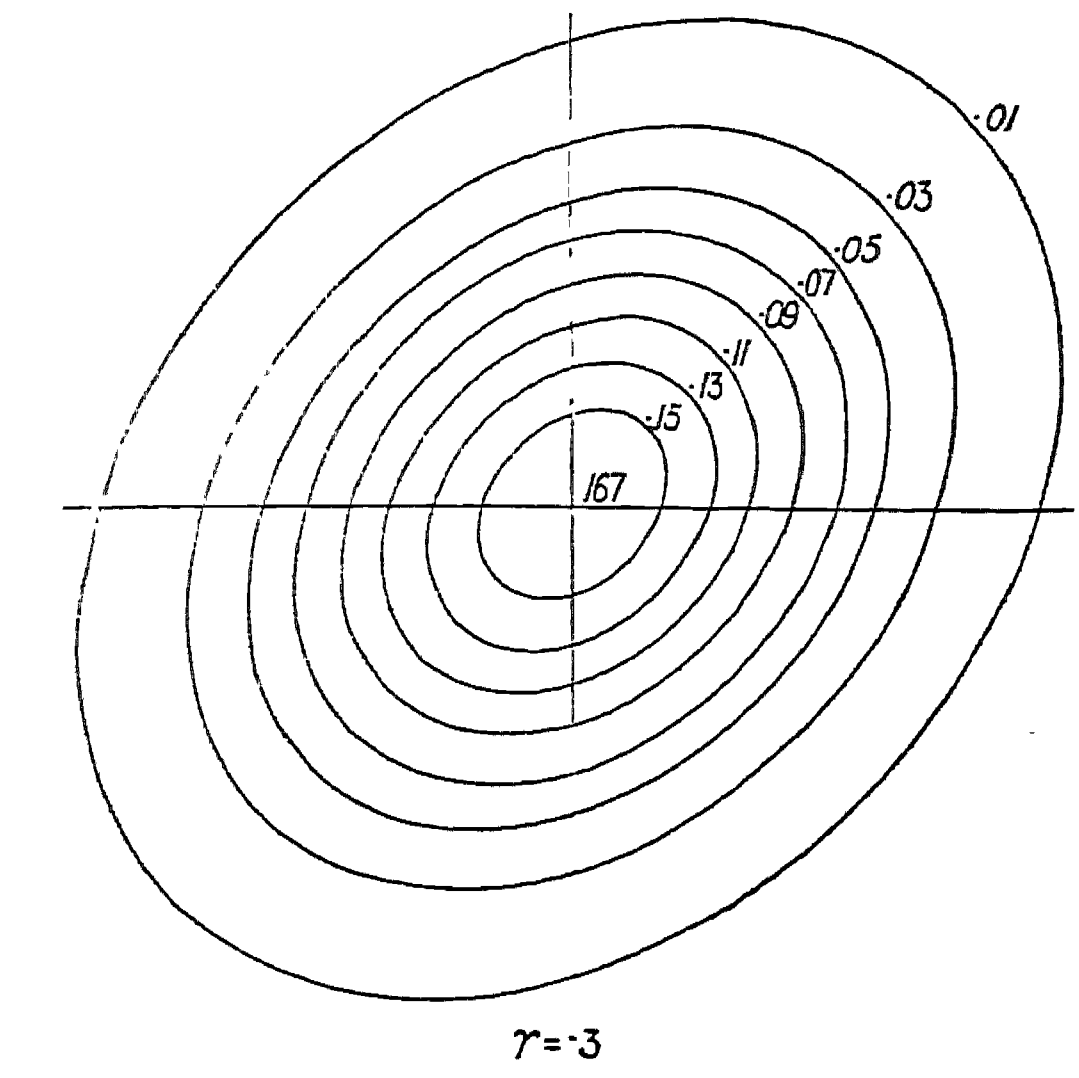


FIG. 9.—NORMAL FREQUENCY SURFACES—*continued.*

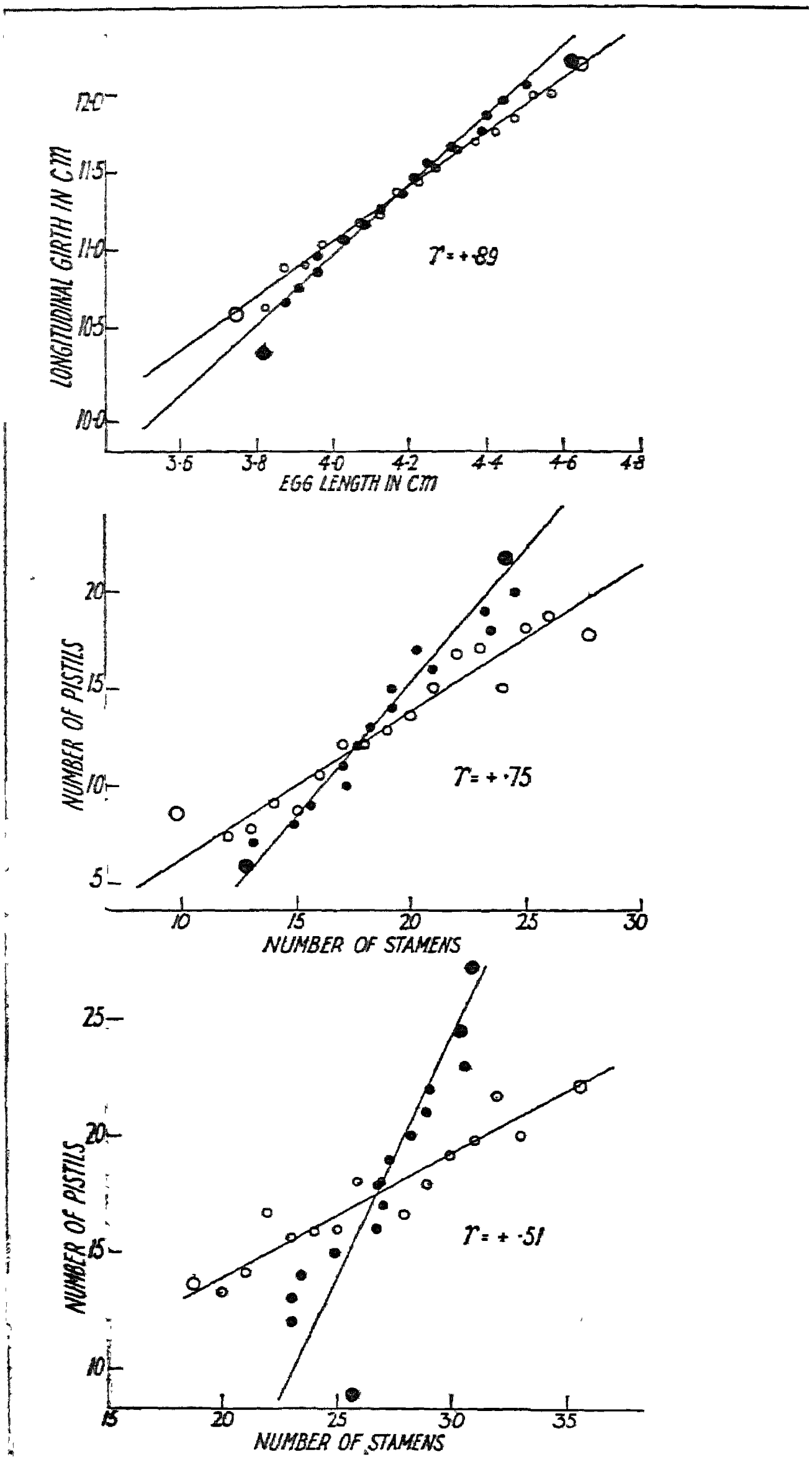


FIG. 10.—REGRESSION DIAGRAMS.

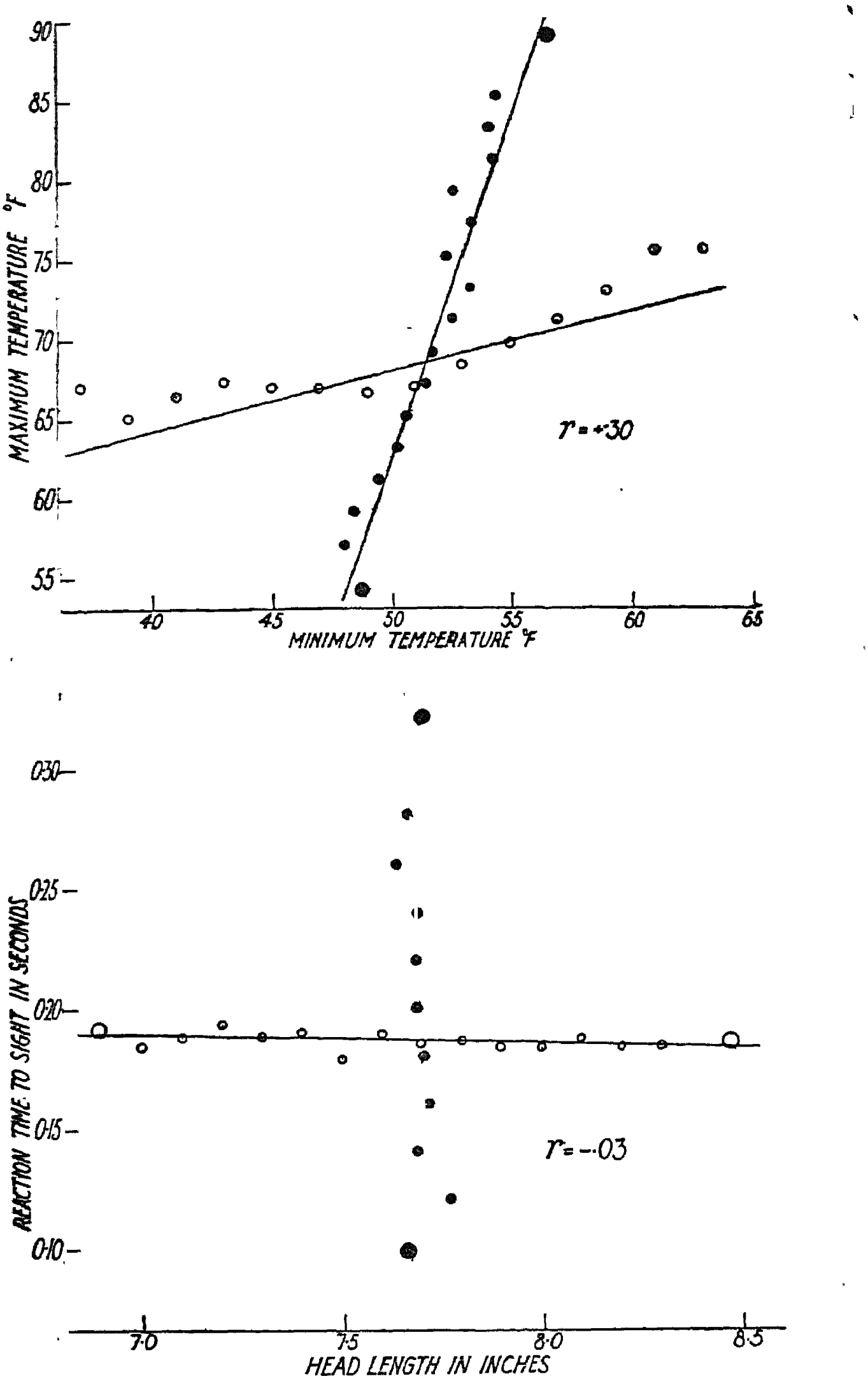


FIG. 10.—REGRESSION DIAGRAMS—*continued.*

the five correlation tables; the means of y for arrays of x being represented by circles, and those of x for arrays of y by dots; in the former x is called the *independent* and y the *dependent* variable, and in the latter, *vice versa*.* The two sets of means are rather irregular, but it is presumed that this is due to the errors of random sampling, and that smooth curves drawn through the points represent more nearly the relationships that would be obtained if the size of the sample were increased indefinitely. The simplest curve that can be drawn, and one which is often a sufficient approximation, is the straight line, and appropriate ones are drawn on Fig. 10. For the maximum and minimum temperatures the circles giving the mean y for values of x lie above the straight line at both extremes,

TABLE 7.11

Length	Number in Group	Mean Girth
3·575	1	9·85
3·725	2	10·70
3·775	6	10·67
Means 3·742	—	10·59

suggesting that the straight line is not quite adequate; however, this is typical of the sort of data to which this method of correlation is often applied as an approximation, and we will use the straight lines. The two sets of means give two straight lines, which become more divergent as the association decreases; indeed, the diagram corresponding to Table 7.5 shows the two lines to be practically perpendicular. The line which is least inclined to the x-axis has x as the

* The means of the extreme groups are naturally very inaccurate, being based on very few observations, and it is usual to combine several such groups. For example, in Table 7.1, if we assume the central values of lengths to be 3·575, 3·625, etc., cm., and those of girths to be 9·85, 9·95, etc., cm., we have the means in Table 7.11, with length as the independent variable. To combine the three readings, we find the weighted means of the length and girth. That of the length is

$$\frac{3{\cdot}575 + 2 \times 3{\cdot}725 + 6 \times 3{\cdot}775}{9} = 3{\cdot}742,$$

and similarly that of the girth is 10·59. Points obtained in this way are shown by larger circles and dots in Fig. 10.

independent variable, and similarly the other, being more nearly parallel to the y-axis, has y as the independent variable.

The general equation of a straight line with x as the independent variable is

$$Y = ax + b$$

where a and b are constants and the capital letter Y denotes the value of the dependent variable given by the line, as distinguished from the actual value of an individual, denoted by y. The problem of estimating the most appropriate values of a and b to fit a line to any given data is analogous to that of determining estimates of parameters for fitting frequency distributions, and the general method of solution that is most in accordance with statistical theory is the *method of least squares*. This consists in choosing the constants a and b so that the sum over all individuals of the quantities $(y - Y)^2$* is reduced to a minimum. A similar line may be obtained giving X in terms of y. The equations to the two lines are deduced in section 7.6 and may be written

$$(Y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad . \quad . \quad . \quad . \quad (7.1)$$

$$(X - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad . \quad . \quad . \quad . \quad (7.11)$$

where $\bar{x}$ and $\bar{y}$ are the two grand means, σ_x and σ_y are the two standard deviations, and r is the correlation coefficient. The value of Y given by equation (7.1) is an estimate of the mean value of the array corresponding to a small sub-range about any given value of x, and that of X given by equation (7.11) is similarly the mean of the array corresponding to a given value of y. Equation (7.1) represents the line that is more nearly parallel to the x-axis; both lines pass through the point $(\bar{x}, \bar{y})$. These are called *regression lines* and the constants

$$r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad r\frac{\sigma_x}{\sigma_y}$$

* In a diagram, these are the distances of the points from the line measured in a direction parallel to the y-axis. They are not perpendicular distances from the line. Further, these points represent individuals, and not array means. If array means are used, a weighted sum of $(\bar{y}_s - Y)^2$ is reduced to a minimum and a similar result is obtained.

are called *regression coefficients*. The coefficients are the tangents of the angles the two lines make with the x- and y-axes respectively.

If $r = 0$, equation (7.1) gives $(Y - \bar{y}) = 0$, an equation which is satisfied by a line parallel to the x-axis, and going through the mean, $\bar{y}$; similarly equation (7.11) gives $(X - \bar{x}) = 0$, and this is satisfied by a line through the mean, $\bar{x}$, and parallel to the y-axis; the two lines are perpendicular. The meaning of these two lines is that the mean value of y is the same for all values of x, and that the mean of x does not change with y, or in other words, that there is no association between x and y; and thus we see that r is zero for zero association or independence. If $r = 1$, one regression coefficient is equal to the inverse of the other, and the two lines coincide. This is the condition for perfect association, i.e. when x is uniquely determined by y (and *vice versa*) and is the state of affairs at which the physicist aims when he controls his experiment so that only the two factors vary and there are no extraneous disturbing influences. The biologist usually has to deal with materials subject to variations over which his control is limited, and consequently he obtains the kind of result illustrated in Tables 7.1 to 7.5, where varying strengths of relationship are shown. If these tables are studied with Fig. 10, it will be noticed that the stronger the relationship, the greater is r and the closer together are the two regression lines. If $\sigma_x = \sigma_y$, or if the diagrams are plotted on such a scale that units of

$$\frac{x - \bar{x}}{\sigma_x} \quad \text{and} \quad \frac{y - \bar{y}}{\sigma_y}$$

are equal, the tangent of the angle between each regression line and the axis of the corresponding independent variable becomes equal to r. If r is positive, the slopes of the lines are in the direction indicating that x increases with y, while if r is negative the slopes are in the opposite direction, showing that x increases as y decreases.

Analysis of Variance

7.23. The preceding section, dealing with regression, pays attention to the relationship between x and y, regarding the deviations rather as errors due to uncontrollable factors; we will now consider the correlation table from the point of view of the variability. Adopting the method of the last chapter, we can analyse the variance of say

egg girth (Table 7.1) into two portions, one associated with differences between the array means for the various groups of length, and the other with residual deviations within the arrays. This can be done by finding the individual array means, as was done in the last chapter, and calculating their variance and the variance of the residual deviations from them. For Table 7.1 we should combine the first five and the last two, giving 18 arrays, or 17 degrees of freedom on which to estimate the variance between arrays, and 937 on which to estimate the residual. If, however, we are satisfied that a straight regression line like equation (7.1) sufficiently represents the trend, the value Y of the girth given by this line for various central values of the length (x) may be used instead of the array means $\bar{y}_s$. For Table 7.1 the line of regression of y on x (as computed in section 7.5) is

$$(Y - 11{\cdot}378) = 1{\cdot}756\,(x - 4{\cdot}190);$$

when $x = 3{\cdot}575$, $Y - 11{\cdot}378 = -1{\cdot}080$ and $Y = 10{\cdot}298$,

when $x = 3{\cdot}725$, $Y - 11{\cdot}378 = -0{\cdot}817$ and $Y = 10{\cdot}561$,

and so on. For the sum of squares of the deviations of the array means from the grand mean (11·378) we take $1 \times (-1{\cdot}080)^2 + 2 \times (-0{\cdot}817)^2$, etc., adding these quantities for all arrays; for the squares of the residual deviations from the array means we take $(9{\cdot}85 - 10{\cdot}298)^2 + (10{\cdot}35 - 10{\cdot}561)^2 + (11{\cdot}05 - 10{\cdot}561)^2$, etc., adding these quantities for all cells in the table, while for the total sum of squares we use the marginal totals as usual. These three sums may be found and entered in a table of analysis of variance in just the same way as if we had used the actual array means. This process can be carried out explicitly as indicated, and the reader is recommended to do it as an exercise; alternatively, all the various sums of squares may be found from some of the constants of the table as shown below. We will set out the process algebraically.

Repeating equation (6.6), section 6.4, except that the variable is changed to y, we have for the first method of analysis,

$$\underset{s}{S}S'(y - \bar{y})^2 = \underset{s}{S}S'(y - \bar{y}_s)^2 + \underset{s}{S}n_s(\bar{y}_s - \bar{y})^2, \quad . \quad . \quad (7.2)$$

and for the second method, using the Y given by equation (7.1) instead of $\bar{y}_s$, we have

$$S(y - \bar{y})^2 = S(y - Y)^2 + S(Y - \bar{y})^2, \quad . \quad . \quad . \quad (7.3)$$

where $S = \underset{s}{S}S'$ is the summation over all individuals. This may be expressed in words—the sum of squares of deviations from the grand mean equals the sum of squares of deviations from the regression line plus the sum (over all observations) of squares of deviations

TABLE 7.6

ANALYSIS OF VARIANCE OF y (EGG GIRTH)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Straight regression line	$S(Y-\bar{y})^2 = r^2S(y-\bar{y})^2$	1	${}_yv_s$
Residual	$S(y-Y)^2 = (1-r^2)S(y-\bar{y})^2$	$N-2$	${}_yv_r$
Total	$S(y-\bar{y})^2$	$N-1$	${}_yv_T$

of the regression values from the grand mean. The terms of equation (7.3) are entered in Table 7.6, and it is shown in section 7.6 that the sums of squares are severally equal to those terms in the table which involve r^2. The constant r is again the correlation coefficient.

TABLE 7.61

ANALYSIS OF VARIANCE OF x (EGG LENGTH)

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Straight regression line	$S(X-\bar{x})^2 = r^2S(x-\bar{x})^2$	1	${}_xv_s$
Residual	$S(x-X)^2 = (1-r^2)S(x-\bar{x})^2$	$N-2$	${}_xv_r$
Total	$S(x-\bar{x})^2$	$N-1$	${}_xv_T$

If the reader has found the sums of squares directly as suggested in the last paragraph, he will also be able to check these relations. Similarly, by using the regression line of equation (7.11) we may analyse the variance of x and arrive at the sums of squares in Table 7.61.

We shall now discuss the degrees of freedom. We have seen that in an infinite population for which x and y are independent, the regression of y on x is a line through the mean, $\bar{y}$, and parallel to the x-axis; it follows that for such a case $(y - Y)$ equals $(y - \bar{y})$ and the residual sum of squares equals the total. For a finite sample, however, owing to random errors a regression line will have a slight slope and the residual sum of squares will be different from the total. Moreover, since the regression line is obtained by the method of *least* squares the residual must always be less than the total. It follows from a proof given by Fisher (1925*b*) that if the total sum of squares has n degrees of freedom, the amount of this difference is, on the average, one nth of the total sum of squares, and that the residual sum of squares divided by $n - 1$ is an unbiassed estimate of the variance of y in the population, having the sampling distribution of an estimate based on $n - 1$ degrees of freedom.* Applying this result to Table 7.6 and remembering that there are $N - 1$ degrees for the total sum of squares, we see that when the association in the population is zero, the three variances in the last column are estimates of the same variance, that of y in the population.

Thus, for zero association in the population, Table 7.6 is analogous to Table 6.3 with two arrays (one degree of freedom). If, now, there is an association between x and y, the difference between the residual and total sum of squares is greater than the proportion due to one degree, ${}_yv_r$ is less than ${}_yv_T$ and ${}_yv_s$ is greater than either. We may carry the analogy between Tables 6.3 and 7.6 further by regarding ${}_yv_s$ as the variance due to the regression line; we shall make no attempt, however, to find an analogue to σ_s^2 in Table 6.3. The value ${}_yv_r$ is an estimate of the true variance within arrays, and also of the variance of the distribution represented by a cross-section of

* This is a special case of a more general result that if to a series of N independent observations an equation or series of equations involving u constants altogether is fitted by the method of least squares, the sum of squares of the deviations from the values given by the equations, i.e. the residual sum of squares, divided by $N - u$ is an unbiassed estimate of the variance in the population. Thus, we may say that each constant so determined from the data absorbs one degree of freedom. The one degree of freedom associated with the linear regression line is due to the slope, the other constant, $\bar{y}$, is eliminated from both the residual and total sum of squares. This result also follows from a proof given by Irwin (1931).

This property is a powerful recommendation for the use of the method of least squares in statistical work.

the frequency surface in a vertical plane parallel to the y-axis. Similar arguments may be applied to the analysis in Table 7.61.

It is thus seen that the two straight lines found by least squares are those which make the residual variances a minimum (or make the variance associated with themselves a maximum), and r is a measure of the amount of the total variation in each quantity associated with the appropriate line. If $r = 1$, the residual variance is zero and all the variation in one quantity is explained by variation in the other, while if $r = 0$, the residual variance is as great as the total, and there is no association between x and y. This point of view (since it uses r^2) takes no account of the sign of r, but is merely concerned with measuring the strength of association without troubling whether y increases with x or whether it increases as x decreases.

General Discussion

7.3. The correlation coefficient is much used as a measure of the strength of association, but it is not easy for the beginner to appreciate its scale of values. It may vary between plus and minus unity, but the sign merely shows the direction of the trend (whether x increases with y or whether one increases as the other decreases), and so as a measure of association we are only concerned with values between 0 and 1; a correlation of $-0{\cdot}5$ is as good as one of $+0{\cdot}5$. If we consider samples in which N is so large that for all practical purposes $N - 1 = N - 2$, we see from Tables 7.6 and 7.61 that v_r/v_T, the ratio of the residual to the total variance, is practically equal to $(1 - r^2)$, and that the ratio of the corresponding standard deviations is $\sqrt{1 - r^2}$. These quantities are given in Table 7.7, and the meaning of any value of the correlation coefficient between two quantities is that if we keep one quantity constant, the variability of the other is reduced in the ratio given in the third column of the table. The scale is very uneven, since a correlation of $0{\cdot}6$ reduces the residual standard deviation to $0{\cdot}8$ of the total, while even if the correlation is as high as $0{\cdot}9$, the residual deviations have a standard deviation of $0{\cdot}436$ of the total. This unevenness of scale is shown by Tables 7.1 to 7.5; a far greater change in association is noticeable between Tables 7.1 and 7.3 for a smaller change in correlation coefficient than between Tables 7.3 and 7.5; i.e. a coefficient of $1{\cdot}0$ is much more than twice as good as one of $0{\cdot}5$. Attempts are sometimes made to explain correlation in terms of chances, and to state that if the coefficient for two characters is $0{\cdot}6$ (say), and knowing one

of them we use the regression formula to guess at the other, we shall be right six times in ten trials. Such an explanation is nonsense, and the truth is that the standard error of our guess will be reduced to $\sqrt{1 - 0{\cdot}6^2} = 0{\cdot}8$ of what it would have been had we not used the formula. Such a reduction in standard deviation may be further interpreted in terms of reduction in frequencies of estimates deviating from the actual values by given amounts, on referring to section 2.52.

The following typical cases may assist the reader. Yule (1927) gives the correlation coefficient between the ages of husbands and wives in England and Wales as 0·91, and that between the age and the standard of elementary school boys is of the same order (Jones,

TABLE 7.7*

Correlation Coefficient r_{12}	Ratio of Variances Residual/Total	Ratio of Standard Deviations σ_r/σ_T	z'
zero	1·00	1·000	zero
0·2	0·96	0·980	0·20
0·4	0·84	0·917	0·42
0·6	0·64	0·800	0·69
0·7	0·51	0·714	0·87
0·8	0·36	0·600	1·10
0·9	0·19	0·436	1·47
1·0	zero	zero	infinity

* The column under z' will be referred to in the next chapter.

1910). The likeness between parents and children is expressed by a coefficient of about 0·47 as far as height and several other physical characters are concerned (Snow, 1911), and that between cousins or uncles and nephews, which is so small as scarcely to be noticeable to the "man in the street," has a correlation coefficient of about 0·26. Similarly, taking periods of six days as units, the correlation between the rainfall and hours of bright sunshine in Hertfordshire is negative (i.e. increased rainfall is associated with decreased sunshine) and about 0·2.

The uses (and abuses) of the coefficient of correlation are many, but it may be always safely regarded as a constant descriptive of the properties of the sample. For instance, Weldon found that the correlation between number of pistils and stamens of late flowers of Table 7.2 was 0·75, while for the early ones, Table 7.3, it was 0·51;

that change in the strength of relationship is as much a characteristic of the flowers as a change (say) in mean height of the plants. Fisher and Hoblyn (1928) give tables to show that the correlation between maximum and minimum temperature is about 0·71 in January and 0·30 in August; this change is gradual from month to month and is a quality of the climate.

If a linear regression formula is being used to predict one character from another, the correlation coefficient may safely be used to define the accuracy of the prediction as has been shown, the standard deviation of the differences between actual and predicted values being $\sqrt{1 - r^2}$ times the standard deviation of the dependent variable.

A rather more dangerous use of this constant, however, is as evidence of causation. If two quantities are associated (as shown by the correlation coefficient) the inference often made is that one is a cause and the other an effect. Such an inference is often erroneous when dealing with quantities susceptible to such close control that the correlation coefficient is unity, but it is particularly unsafe when there are uncontrolled variations and the relationship is not exact. Often two quantities are both affected in the same way by a third so that they appear to be related, when actually neither if altered independently would have any effect on the other. For instance, Yule refers to the fact that the proportion of marriages contracted outside the Church of England has for many years been increasing, while the average age at death has also been increasing, and there is a positive correlation; but no one supposes that there is a causal relationship and that a law prohibiting the solemnisation of marriages in churches would have the effect of improving the longevity of the nation. When investigating causation, it is usually well to decide first on other grounds that a causal relationship between two factors is likely, and then to conduct a close analysis of other possible factors before using the correlation coefficient as evidence. Care, common sense, imagination, and a technical knowledge of the subject to which it is applied are particularly necessary in this use of correlation. We shall in Chapter XI discuss methods of analysing data for, and eliminating other factors.

When considered as the expression of the relationship between two quantities (the second method of approach above), the correlation coefficient merely measures the importance of the variations in one quantity associated with the other (the independent variable)

relative to all other variations, and if the variance of the independent variable can be controlled, either experimentally or by selection, the correlation can be altered. It is thus not a physical quantity in the way the regression is. To demonstrate this, we will suppose the independent variable to be x, the dependent one y, the regression a, the correlation coefficient r, and the standard deviations σ_x and σ_y. Then if σ_r is the standard deviation of the residual deviations of y from the regression line, which we may suppose to be constant for all values of x,

$$\sigma_r = \sigma_y \sqrt{1 - r^2}.$$

But

$$a = r\frac{\sigma_y}{\sigma_x} = \frac{r}{\sqrt{1 - r^2}}\frac{\sigma_r}{\sigma_x},$$

whence

$$r^2 = \frac{a^2{\sigma_x}^2}{{\sigma_r}^2 + a^2{\sigma_x}^2}.$$

Now σ_r and a are assumed to be constant, so that r is not independent of σ_x. For the relationship between numbers of pistils x and stamens y, $r = 0{\cdot}749$; then $\sigma_x = 3{\cdot}388$, $\sigma_y = 3{\cdot}298$, $\sigma_r = 2{\cdot}185$, $a = 0{\cdot}729$ and $a\sigma_x = 2{\cdot}470$. If now we had selected flowers so that the standard deviation of the number of pistils had been reduced by one-half, we should have reduced the correlation coefficient from $0{\cdot}749$ to $0{\cdot}492$, and we should have increased it to $0{\cdot}914$ if we had been able to increase the variation of pistils so as to double the standard deviation (these results may be obtained by substitution in the above formula). The same consideration applies when the effect of some factor like temperature on some variable quantity is being investigated experimentally; by altering the range of temperatures, the correlation coefficient may be altered considerably, and so it is a result of the arbitrary experimental conditions. When such conditions are under the control of the experimenter, the amount of variation in x should be made as great as possible so as to make r large; we shall see in the next chapter that this gives a maximum precision to the measurement of a. A low correlation coefficient does not necessarily mean that x is incapable of having an important influence on y; it may mean that an insufficiently large range of x has been tried. Or if x cannot be varied, it may be that although

its effect on y as measured by the correlation coefficient is small relative to disturbing factors, its absolute effect as measured by the regression coefficient is important. The correlation coefficient between the yield of some crop over several plots and quantities of fertiliser used may be low, but if the regression is appreciable, the total advantage of the use of fertiliser by the whole agricultural industry will be considerable. In such instances, except as an indication of the existence of association, the coefficient of correlation is an unsuitable constant, and that of regression is more suitable.

Those whose previous outlook has been physical often prefer to regard the scatter of points on a scatter diagram as being due to disturbing factors akin to experimental errors, and as obscuring an exact physical relationship which is assumed to underlie the data. To such people the regression obtained by the method of least squares appears to represent the best possible approach to the true relationships, and when they realise that there are two regression lines, they are puzzled to know which to take. Such an interpretation of a scatter diagram or correlation table may be a little dubious, and in any case, without additional knowledge, it is not possible to arrive at a theoretically satisfactory unique line. It is safer to regard all the features of the table—including the scatter—as equally real properties of the sample, and to regard the means, variances, correlations and regressions as so many constants descriptive of some of those properties.

To sum up; there are three important constants that express the properties of a correlation table:

(1) The correlation coefficient that measures the importance of the variation in y associated with x relative to the total variation,
(2) The regression coefficient that measures the average amount of increase or decrease in y per unit increase in x, and
(3) The residual variance that measures the scatter of values of y about the regression line.

For different populations, these constants are independent in that a high value of one constant does not necessarily mean a high or low value of either of the others.

Estimation of Coefficients

7.4. The method of maximum likelihood may be applied to determining estimates of the parameters of the normal frequency surface

of section 7.21. For $\bar{x}$, $\bar{y}$, σ_x and σ_y it gives the estimates already deduced for single distributions and for r it gives

$$r = \frac{S(x - \bar{x})(y - \bar{y})}{\sqrt{S(x - \bar{x})^2 S(y - \bar{y})^2}}, \quad . \quad . \quad . \quad . \quad (7.4)$$

where S is the summation over all individuals.

To be consistent with the earlier treatment of frequency distributions we should denote the population value by ρ and describe the above value of r as an estimate obtained from a sample.

From equations (7.1), (7.11) and (7.4) we have for the regression coefficients

$$\textit{Regression of } y \textit{ on } x = r\frac{\sigma_y}{\sigma_x} = \frac{S(x - \bar{x})(y - \bar{y})}{S(x - \bar{x})^2} \quad . \quad (7.5)$$

and

$$\textit{Regression of } x \textit{ on } y = r\frac{\sigma_x}{\sigma_y} = \frac{S(x - \bar{x})(y - \bar{y})}{S(y - \bar{y})^2} \quad . \quad (7.51)$$

If N is the size of the sample,

$$\frac{S(x - \bar{x})(y - \bar{y})}{N}$$

is called the first product moment ($= p$, say), whence we see that

$$r = \frac{N}{(N - 1)} \frac{p}{\sigma_x \sigma_y} \quad . \quad . \quad . \quad . \quad . \quad (7.41)$$

where σ_x and σ_y are calculated correctly on $N - 1$ degrees of freedom; if the sample is large $N/(N - 1)$ may be taken as nearly equal to unity.

For moderately small samples, r may be calculated directly from the individual observations, using equation (7.5), but when the sample is large and the data are grouped, a rather more elaborate method of computation, such as is illustrated in the next section, is advisable.

Computation of Coefficients

7.5. The denominators of the above equations may be computed by the methods of section 1.31 or 6.3, and here we need only concern ourselves with the product term in the numerators. It will be con-

venient to use similar methods. Corresponding to equation (1.2), p. 37, we have

$$p = \frac{Sxy}{N} - \bar{x}\bar{y} \quad . \quad . \quad . \quad . \quad . \quad . \quad (7.6)$$

and corresponding to equation (6.5), p. 131, we have

$$S(x - \bar{x})(y - \bar{y}) = Sxy - N\bar{x}\bar{y} = Sxy - \frac{T_x T_y}{N} \quad . \quad . \quad (7.61)$$

where x and y may be measured from any convenient origin in the true units. The correcting term in the above equations may be positive or negative, and if either $\bar{x}$ or $\bar{y}$ is zero, this term is zero.

When the observations are grouped, all those in any group may be assumed to be at its centre and it is convenient to measure the deviations from a group near the middle of each distribution in terms of the sub-range h_x or h_y. Then if p' is the product moment in such units,

$$p = h_x h_y p' . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7.62)$$

Sheppard's corrections may be applied in calculating the denominators of equations (7.4) and (7.5) from grouped data; there is no corresponding correction for the numerator.

We shall compute the constants for Table 7.1 as an example, where the arbitrary values x' and y' are given in the column or row adjacent to that containing the true values of the variate. The computations are all contained in Table 7.8, where columns (1), (2), (3), (4), (7), (8), (9) and (10) contain all the data for obtaining the means and the second moments, and below the table all the calculations have been carried out as in the first chapter. Sheppard's corrections have been applied, and as the sample is a large one, the second moment has been obtained by dividing the sums of squares by the number of observations (995) instead of by the degrees of freedom.

To find the product moment, we must first find $Sx'y'$. This could be done directly by writing in the corner of each cell in Table 7.1 its $x'y'$ (the one in the top left-hand corner would be $-14 \times -11 = +154$, the next in the top row would be $-14 \times -10 = +140$, and so on), and then multiplying each $x'y'$ by the number in the cell and adding; in this table, all the squares in the two quadrants containing most observations would have positive products, and those

in the other two would have negative ones. A better method, however, is to do the summation in two parts, keeping y' constant first, summing all the x''s in each array of y' separately, and then adding the arrays. If we consider any array of y' with n_y observations:

$$Sx'y' = \underset{s}{S}y'S'x'_y = \underset{s}{S}y'n_y\bar{x}'_y, \qquad . \quad . \quad . \quad (7.63)$$

where S, $\underset{s}{S}$ and S' are the summations used previously, and $\bar{x}'_y$ is the mean x' of all the observations in the sth array of y'.

The process is carried out in columns (5) and (6) of Table 7.8. In the array $y' = -14$, $\bar{x}'_y$ is $-11/1$, so $n_y\bar{x}' = -11$; $y' = -9$, $\bar{x}'_y = -12/2$, so $n_y\bar{x}'_y = -12$, and so on for the other quantities in column (5); the sum of this column is the sum of all the deviations of x' from the arbitrary origin, and so should equal the sum of column (9). The terms of column (6) are the products of those in columns (1) and (5), and their sum is the required $Sx'y'$. As a check on the arithmetic, this may be done the other way, summing first for each separate array of x' and adding them. This is done in columns (11) and (12), and if the arithmetic is correct, the totals of columns (11) and (3) and those of columns (6) and (12) should be equal. There is no independent check of columns (4) and (10), and so they should be repeated carefully.

The product moment is calculated below Table 7.8 by applying equation (7.6) to the arbitrary values and then correcting by equation (7.62). In finding r and the regressions, we have used p, σ_x and σ_y in the natural units of centimetres, but it would be just as easy to maintain arbitrary units throughout and then to correct the regressions by multiplying by h_y/h_x and h_x/h_y. The correlation coefficient, being a pure number, needs no such correction.

Regression Lines and Sums of Squares: Proofs

7.6. Let x and y be the individual values of the variate with means $\bar{x}$ and $\bar{y}$; then it is required to find the constants a and b of the following equation:

$$Y = ax + b, \qquad . \quad . \quad . \quad . \quad . \quad (7.7)$$

such that $S(y - Y)^2$ is a minimum.
From (7.7)

$$S(y - Y)^2 = S(y - ax - b)^2,$$

TABLE 7.8

(1) y'	(2) n_y	(3) $n_y y'$	(4) $n_y y'^2$	(5) $n_y \bar{x}'_y$	(6) $n_y \bar{x}'_y y'$
− 14	1	− 14	196	− 11	154
− 13	—	—	—	—	—
− 12	—	—	—	—	—
− 11	—	—	—	—	—
− 10	—	—	—	—	—
− 9	2	− 18	162	− 12	108
− 8	2	− 16	128	− 12	96
− 7	6	− 42	294	− 32	224
− 6	15	− 90	540	− 75	450
− 5	10	− 50	250	− 43	215
− 4	35	− 140	560	− 117	468
− 3	38	− 114	342	− 110	330
− 2	79	− 158	316	− 141	282
− 1	90	− 90	90	− 65	65
0	119	—	—	+ 9	—
1	114	+ 114	114	141	141
2	101	202	404	185	370
3	98	294	882	284	852
4	76	304	1 216	285	1 140
5	66	330	1 650	283	1 415
6	48	288	1 728	265	1 590
7	29	203	1 421	186	1 302
8	16	128	1 024	122	976
9	6	54	486	60	540
10	3	30	300	28	280
11	1	11	121	11	121
Totals ..	955	+ 1 226	12 224	+ 1 241	+ 11 119

$$y\nu'_1 = \bar{y}' = \frac{+\ 1\ 226}{955} = +\ 1{\cdot}283\ 77, \qquad h_y = 0{\cdot}1 \text{ cm.}$$

$$\text{mean girth} = 11{\cdot}25 + 0{\cdot}128 = 11{\cdot}378 \text{ cm.}$$

$$y\nu'_2 = \frac{12\ 224}{955} = 12{\cdot}800\ 0$$

$$-\ y\nu_1'^2 = -\ 1{\cdot}648\ 1$$

$$y\nu_2 = 11{\cdot}151\ 9$$

$$-\ 0{\cdot}083\ 3$$

$$y\mu_2 = 11{\cdot}068\ 6$$

$$\sigma_y^2 = h_y^2 \times 11{\cdot}068\ 6 = 0{\cdot}110\ 69 \text{ cm.}^2$$

$$\sigma_y = 0{\cdot}332\ 70 \text{ cm.}$$

$$p' = \frac{+\ 11\ 119}{955} = 11{\cdot}642\ 9$$

$$-\ x\nu'_1 y\nu'_1 = \frac{-\ 1{\cdot}668\ 2}{+\ 9{\cdot}974\ 7}$$

$$p = h_x h_y \times 9{\cdot}974\ 7 \text{ cm.}^2 = +\ 0{\cdot}049\ 874 \text{ cm.}^2$$

$$r = +\ \frac{0{\cdot}049\ 874}{0{\cdot}332\ 70 \times 0{\cdot}168\ 54} = +0{\cdot}889$$

TABLE 7.8—*continued*

(7) x'	(8) n_x	(9) $n_x x'$	(10) $n_x x'^2$	(11) $n_x \bar{y}'_x$	(12) $n_x \bar{y}'_x x'$
−11	1	− 11	121	− 14	154
−10	—	—	—	—	—
− 9	—	—	—	—	—
− 8	2	− 16	128	− 11	88
− 7	6	− 42	294	− 35	245
− 6	14	− 84	504	− 75	450
− 5	16	− 80	400	− 59	295
− 4	30	−120	480	−105	420
− 3	55	−165	495	−126	378
− 2	69	−138	276	−119	238
− 1	85	− 85	85	− 70	70
0	128	—	—	+ 25	—
1	101	+101	101	106	106
2	98	196	392	185	370
3	91	273	819	240	720
4	106	424	1 696	414	1 656
5	54	270	1 350	241	1 205
6	40	240	1 440	215	1 290
7	25	175	1 225	147	1 029
8	15	120	960	111	888
9	12	108	972	91	819
10	2	20	200	17	170
11	5	55	605	48	528
Totals	955	+1 241	12 543	+1 226	+11 119

$$ {}_x\nu'_1 = \bar{x}' = \frac{+\ 1\,241}{955} = +\ 1{\cdot}299\,48, \qquad h_x = 0{\cdot}05 \text{ cm.} $$

mean length = 4·125 + 0·065 = 4·190 cm.

$$ {}_x\nu_2' = \frac{12\,543}{955} = \quad 13{\cdot}134\,0 $$

$$ \begin{aligned} -\ {}_x\nu'^2_1 &= -\ 1{\cdot}688\,6 \\ {}_x\nu_2 &= \quad 11{\cdot}445\,4 \\ &\quad -\ 0{\cdot}083\,3 \\ {}_x\mu_2 &= \quad 11{\cdot}362\,1 \end{aligned} $$

$$ \sigma_x^2 = h_x^2 \times 11{\cdot}362\,1 = 0{\cdot}028\,405 \text{ cm.}^2 $$

$$ \sigma_x = 0{\cdot}168\,54 \text{ cm.} $$

Regressions :—

$$ y \text{ on } x = +\ \frac{0{\cdot}049\,874}{0{\cdot}028\,405} = +\ 1{\cdot}756 \text{ cm./cm.} $$

$$ x \text{ on } y = +\ \frac{0{\cdot}049\,874}{0{\cdot}110\,69} = +\ 0{\cdot}450\,6 \text{ cm./cm.} $$

and differentiating this with respect to a and b respectively, and equating to zero (the usual method of evaluating constants to reduce a function of them to a minimum), we have

$$\begin{aligned} -2Sx(y-ax-b) &= 0, \\ -2S(y-ax-b) &= 0; \end{aligned}$$

or expanding term by term, and cancelling out common terms,

$$\left.\begin{aligned} Sxy &= aSx^2 + bSx, \\ Sy &= aSx + bS1. \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (7.8)$$

Now $Sx = N\bar{x}$, $Sy = N\bar{y}$ and $S1 = N$, and using these we may solve (7.8) to obtain

$$a = \frac{Sxy - N\bar{x}\bar{y}}{Sx^2 - N\bar{x}^2}, \quad . \quad . \quad . \quad . \quad . \quad (7.81)$$

and

$$b = \bar{y} - a\bar{x}.$$

We have seen that $Sxy - N\bar{x}\bar{y} = S(x-\bar{x})(y-\bar{y})$ and $Sx^2 - N\bar{x}^2 = S(x-\bar{x})^2$; using these in (7.81) and substituting in (7.7) we obtain

$$Y - \bar{y} = \frac{S(x-\bar{x})(y-\bar{y})}{S(x-\bar{x})^2} \cdot (x-\bar{x}), \quad . \quad . \quad (7.71)$$

which reduces to regression equation (7.1), when the value for r given in equation (7.4) and the standard deviations are substituted for the sums of products and squares.

Rewriting equation (7.3) by substituting the regression value of equation (7.1) for Y, we have

$$S(y-\bar{y})^2 = S\left\{(y-\bar{y}) - r\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right\}^2 + r^2\frac{\sigma_y{}^2}{\sigma_x{}^2}S(x-\bar{x})^2. \quad (7.9)$$

Now

$$\frac{\sigma_y{}^2}{\sigma_x{}^2} = \frac{S(y-\bar{y})^2}{S(x-\bar{x})^2}$$

and the second term on the right-hand side of equation (7.9) becomes

$$r^2S(y-\bar{y})^2.$$

Also expanding the first term on that side and substituting for r from equation (7.4) we find that the term reduces to

$$(1 - r^2)S(y - \bar{y})^2.$$

Equation (7.9) may thus be written

$$S(y - \bar{y})^2 = (1 - r^2)S(y - \bar{y})^2 + r^2S(y - \bar{y})^2, \quad . \quad (7.91)$$

and these are the sums of squares entered in Table 7.6.

Further, by substituting for r in equation (7.91) we see that the sum of squares of the deviations of y from the regression line is

$$S(y - Y)^2 = S(y - \bar{y})^2 - \frac{[S(x - \bar{x})\,(y - \bar{y})]^2}{S(x - \bar{x})^2} \quad . \quad (7.92)$$

We shall use this in subsequent chapters.

By a similar series of calculations, interchanging x and y, we can obtain the regression line and sums of squares for x.

CHAPTER VIII

SAMPLING ERRORS OF CORRELATION AND REGRESSION CONSTANTS

LIKE all other statistical constants, the correlation coefficient is subject to errors of random sampling, so that different samples from the same population give varying values which themselves form a frequency distribution. In this chapter we shall deal with this distribution and with tests of significance based upon it, and with the sampling errors of regression constants.

POPULATIONS WITH ZERO ASSOCIATION

8.1. One of the commonest tests is to see if an observed correlation coefficient is significantly greater than zero, and to do this, the probability that such a value could arise in a random sample from a population having no association has to be found. For samples from such a population, the standard error of the distribution of r is $\pm 1/\sqrt{N-1}$ (or for large numbers, $\pm 1/\sqrt{N}$ is an adequate approximation) where N is the number of individuals, i.e. the number of pairs of readings of x and y. Unless the sample is very small, the distribution is near enough to normality to justify the use of the criterion that a value of r more than twice the standard error for zero association is above the 0·05 level of significance. For the data of Table 7.5 there are 4 690 observations, and the correlation between reaction time to sight and head length is — 0·034, with a standard error of ± 0·015. The value is a little more than twice its standard error, and so is suggestive of a real, although exceedingly weak, association. It would not be safe, however, to deduce very much from such a result; indeed, a correlation of that sort might conceivably arise if there were a small personal error in measurement such that the observer who overestimated the head-length tended slightly to underestimate the reaction time, and if several observers took the measurements, the same one always measuring both characters on the same individual. For the maximum and minimum temperatures of Table 7.4, $N = 1\,518$, $r = +\,0{\cdot}30$, and being more than ten times its standard error

$$\left(\pm\frac{1}{\sqrt{1\,518}} = \pm\,0{\cdot}026\right)$$

is undoubtedly real. Hence, sufficient observations can establish the reality of an exceedingly weak association, and the probability of its significance is no guide to the closeness of the relationship.

We have seen, however, from Tables 7.6 and 7.61 that correlation may be expressed as an analysis of variance, where one degree of freedom is taken by the regression line and $(N - 2)$ are taken by the residuals. For a sample from a population with zero correlation the two variances ${}_yv_s$ and ${}_yv_r$ are independent estimates of the same variance, and we may test for the significance of their difference by finding

$$z = \tfrac{1}{2} \log_e \frac{{}_yv_s}{{}_yv_r} = \tfrac{1}{2} \log_e \frac{r^2(N-2)}{1-r^2} \quad . \quad . \quad . \quad (8.1)$$

and seeing if it lies beyond the 5 or 1 per cent. point of Fisher's tables for $n_1 = 1$ and $n_2 = N - 2$. Alternatively, according to section 6.5 we may use the t-test, finding

$$t = \sqrt{\frac{r^2(N-2)}{1-r^2}} \quad . \quad . \quad . \quad . \quad . \quad (8.2)$$

and testing its significance for $N - 2$ degrees of freedom.

Fisher (1936) has prepared tables giving values of the correlation coefficient on different levels of significance for samples of various sizes. When there are 12 degrees of freedom ($N = 14$), $r = 0{\cdot}532\,4$ lies on the 0·05 level, and making the above transformation, we find that $z = 0{\cdot}778\,8$ and $t = 2{\cdot}179$. Both of these lie on the 5 per cent. level of their respective distributions when $n_1 = 1$ and $n_2 = 12$ or $n = 12$. We can now check the adequacy of the assumption of the normal distribution of r for moderately large samples; when $N = 20$ (say), $1/\sqrt{N-1} = 0{\cdot}229$, and a correlation coefficient of 0·46 lies on the 0·05 level, while Fisher's tables give the exact value for 18 ($= N - 2$) degrees of freedom as 0·444; when $N = 10$ the corresponding values are $2 \times 0{\cdot}333 = 0{\cdot}67$ for the normal distribution and 0·632 for the true distribution, and at $N = 5$ they are $2 \times 0{\cdot}500 = 1{\cdot}0$ and 0·878. As far as the 0·05 level of significance goes, the assumption of normality is not likely to lead to serious error even for samples as small as 10, but below that it makes the test unnecessarily stringent; the true test based on z or t, however, is correct for samples of all sizes.

Contrary to an opinion often expressed, an association measured by the correlation coefficient may be significant in a very small sample

if it is strong enough; and this is in accordance with common experience. A physicist, for instance, if he obtains on a graph half a dozen

TABLE 8.1

PERCENTAGE PROTEIN CONTENT

Year t	Corn Selected for			
	High Protein Content		Low Protein Content	
	y_1 Mean	v_1 Variability	y_2 Mean	v_2 Variability
1896	10·93	9·50	10·93	9·50
1897	10·99	10·90	10·63	8·47
1898	10·98	11·15	10·49	12·61
1899	11·62	11·00	9·59	10·50
1900	12·62	8·09	9·13	11·34
1901	13·78	8·48	9·63	11·47
1902	12·90	8·50	7·86	9·60
1903	13·51	10·04	8·00	10·41
1904	15·03	9·05	8·17	9·91
1905	14·73	8·55	8·58	9·91
1906	14·26	9·19	8·65	10·64
1907	13·90	10·72	7·32	12·57
1908	13·94	11·91	8·96	14·06
1909	13·29	10·76	7·48	12·57
1910	14·87	9·68	8·26	10·41
1911	13·79	12·98	7·90	14·81
1912	14·49	7·80	8·23	9·96
1913	14·83	8·23	7·71	12·32
1914	15·04	9·44	7·67	12·39
1915	14·54	10·19	7·27	11·69
1916	15·66	8·56	8·68	11·86
1917	14·45	12·80	7·09	10·01
1918	15·49	8·78	7·13	10·52
1919	14·70	10·54	6·46	8·05
1920	14·01	12·78	7·54	11·80
1921	16·66	11·04	9·14	14·77
1922	17·34	7·15	7·42	9·43
1923	16·53	8·50	6·48	11·27
1924	16·60	7·17	8·38	13·96

points which lie anywhere near a straight line does not say that there is no relationship but assumes one and draws his line. This is the

more extreme case in which the association is high, and the correlation coefficient is a device for measuring and testing the association more objectively and exactly, and is particularly useful in border-line cases.

We shall take as example some data by Winter (1929), for which corn was grown for twenty-nine years in two series of plots; in the first, the seed for one year was from ears of corn in the previous year's crop in that series, selected because it had high protein content, and in the second the seed for one year was selected from the previous year's crop because of low protein content. The data given in Table 8.1 are the mean percentages of protein for each year, and the coefficients of variation per cent. from ear to ear, and we will call them y_1, y_2, v_1 and v_2. The mean protein percentages of the first series (y_1) seem to have increased progressively, and those of the second series (y_2) appear to have decreased, while it is difficult to see what tendency the variabilities show; we will investigate these trends by means of the correlation coefficients, thus assuming them to be sufficiently well expressed by a linear relationship with time. The correlation coefficients between time and y_1, v_1, y_2 and v_2 are $+0{\cdot}862$, $-0{\cdot}081$, $-0{\cdot}708$ and $+0{\cdot}256$ respectively.* For 27 degrees of freedom, $r = 0{\cdot}367$ lies on the $0{\cdot}05$ level of significance so that on the average the mean percentages of protein have increased significantly in the first series and decreased in the second with time, but the continued selection has had no measurable effect on the variability.

Populations in which Association is not Zero

8.2. When the association is not zero, the standard error of the correlation coefficient is

$$\frac{1-\rho^2}{\sqrt{N-1}}$$

where ρ is the correlation coefficient in the population; but this is not of much use in testing whether one observed value is greater than another, because (*a*) unless ρ is small, the distribution of r, the correlation coefficient of the sample, is far from normal, and (*b*) the population value (ρ) is unknown, and if the sample value (r) is substituted the error is likely to be serious. These considerations

* The computation is much simplified if the year 1910 is chosen as zero time, and the others are labelled $-1, -2, \ldots, +1, +2, \ldots$; then $\bar{t} = 0$.

become particularly important when ρ approaches 0·8, for the sampling curve becomes exceedingly skew; for evidence of this, the reader is referred to *A Co-operative Study* (1917), in which very full distributions are worked out by K. Pearson and his colleagues from a formula developed by Fisher (1915).*

These difficulties in the use of the sampling distribution of r may be overcome by making a transformation suggested by Fisher (1921), which consists in writing

$$z' = \tanh^{-1} r = \tfrac{1}{2}\{\log_e (1 + r) - \log_e (1 - r)\} \quad . \quad (8.3)$$

This is not quite the same as the z found from the ratio of two variances, so we have added the dash to make a distinction. This quantity has a standard error of

$$\frac{1}{\sqrt{N - 3}}$$

(i.e. it is independent of the value of z'), and although the distribution is not quite normal, it is nearly so, and for most tests, even on small samples, normality may be assumed. Fisher (1936) gives a table relating z' to r, and some of the values are entered in Table 7.7; when $r = 1{\cdot}0$, $z' = \infty$, and the whole effect of the transformation is to give a more open scale for z' when the association is high, when, as we saw in the last chapter, small changes in r correspond to important changes in the variance absorbed by the regression line.

Significance of Differences between Correlations

8.21. Because its distribution is more nearly normal, Fisher recommends the z' transformation when testing the significance of the deviation of an observed correlation from zero; but the effect of the transformation is small for small values of r, and this step is not important. Differences of observed correlations from some assumed population value, or between pairs of observed correlations, may be

* It may be well to mention here a fairly common fallacy in the use of the standard error for testing the significance of an observed coefficient, r. As an approximation, $(1 - r^2)/\sqrt{N - 1}$ is often written for the unknown $(1 - \rho^2)/\sqrt{N - 1}$ and if any observed correlation is greater than twice the former standard error it is judged to be significant; but if the hypothesis being tested is that $\rho = 0$, the true standard error of r is $1/\sqrt{N - 1}$ as shown in the previous section.

tested by using z' and its standard error just as we did the mean and its standard error in large samples.

For example, Snow (1911) gives the correlation coefficient between brothers and sisters, measured for a variety of characters, as 0·521, while Fisher (1928) found from 34 pairs of unlike sex from triplet births the correlation coefficient for cubit to be 0·645; can this difference be attributed to random errors? The values of z' are 0·578 and 0·767, and the standard error of the second is

$$\pm \frac{1}{\sqrt{31}} = \pm 0{\cdot}180$$

(we assume the first determination to be the exact population value, since it is based on many more observations); the difference is less than twice the standard error and is insignificant.

In this example, deviations of z' of $\pm$ 0·360 from the population value (0·578) lie on the 0·05 level of significance, giving values of z' on this level of 0·218 and 0·938; the corresponding values of r are 0·214 and 0·734. Thus the chances are about 20:1 against a random sample of 34 pairs giving values of r outside these limits, but we regard values anywhere between them as common experience; the deviations of the limits of r from 0·521 are — 0·307 and + 0·213, and their inequality shows the skewness of its distribution.

The following is an example of the use of z' to test the significance of the difference between two observed correlations. Table 7.2 shows the correlation between number of stamens and pistils measured on 373 flowers collected late in the season to be 0·745, while a sample of 268 flowers collected earlier gave a correlation coefficient of 0·506; making the transformation, we find $z' = 0{\cdot}962$ and 0·557, and the difference, being 0·405 with a standard error of

$$\sqrt{\frac{1}{370} + \frac{1}{265}} = 0{\cdot}080,$$

is quite significant.

The Combination of Estimates of a Correlation

8.22. It sometimes happens that we have a number of correlation coefficients estimated from several small samples from populations having the same coefficient, and that we wish to combine them to obtain a mean. This is best done by making the transformation,

finding the mean z', and then re-transforming back to r. If the samples are not of equal size, we naturally give the larger ones more weight than the smaller ones; but the weights should be proportional, not to N, the size of the sample, but to $(N - 3)$. Thus if $z'_1, z'_2 \ldots$ are the individual estimates based on samples of $N_1, N_2 \ldots$ pairs of observations, the weighted mean is

$$\bar{z}' = \frac{(N_1 - 3)z'_1 + (N_2 - 3)z'_2 + \ldots}{(N_1 - 3) + (N_2 - 3) + \ldots}$$

and the standard error of $\bar{z}'$ is

$$\frac{1}{\sqrt{(N_1 - 3) + (N_2 - 3) + \ldots}}.*$$

Table 8.2 contains the correlation coefficients between cephalic index and upper face form for samples of skulls belonging to thirteen races (Tschepourkowsky, 1905). The values of z' are in the fourth column, and their mean, when weighted with $(N - 3)$, and standard error are

$$\frac{-42{\cdot}408}{696} = -0{\cdot}060\,9, \quad \text{and} \quad \pm\frac{1}{\sqrt{696}} = \pm 0{\cdot}037\,9.$$

The mean z' is not significant.

If an improved estimate of the correlation is obtained by combining a large number of very small samples in this way, there is a possibility of serious error in $\bar{z}'$ due to the fact that its distribution is not quite normal and there is a slight bias. For this reason, the number of coefficients combined should be small compared with the average size of sample.

The Significance of Groups of Correlations

8.23. If we have a group of correlation coefficients, however, and wish to see if, as a whole, they show association, a test involving z' and χ^2 may be used. Since the standard error of z' is $1/\sqrt{N-3}$, the quantity $z'\sqrt{N-3}$ is distributed approximately normally and

* In the language of section 3.7, $N_1 - 3$, $N_2 - 3$, etc., are the quantities of information given by the samples. In finding the mean of z', these quantities are the weights, and since the quantities are additive, the variance of the mean z' is the inverse of the total quantity of information.

with unit standard deviation in samples from a population with zero association, and we may calculate

$$\chi^2 = S\left(z'\sqrt{N-3}\right)^2,$$

where S is the summation over the number of samples available. As we have indicated in section 4.5, this should be distributed as the χ^2 for g degrees of freedom if there are g samples, and we may find P, the probability that random samples from the

TABLE 8.2

CORRELATION COEFFICIENTS BETWEEN CEPHALIC INDEX AND UPPER FACE INDEX

Race	Number of Cases N	Correlation Coefficient	z'	$z'^2(N-3)$
Australians	66	+ 0·089	+ 0·089	0·50
Negroes	77	+ 0·182	+ 0·184	2·51
Duke of York Islanders	53	− 0·093	− 0·093	0·43
Malays	60	− 0·185	− 0·187	1·99
Fijians	32	+ 0·217	+ 0·221	1·42
Papuans	39	− 0·255	− 0·261	2·45
Polynesians	44	+ 0·002	+ 0·002	0·00
Alfourous	19	− 0·302	− 0·312	1·56
Micronesians ..	32	− 0·251	− 0·257	1·92
Copts	34	− 0·147	− 0·148	0·68
Etruscans	47	− 0·021	− 0·021	0·02
Europeans	80	− 0·198	− 0·201	3·11
Ancient Thebans ..	152	− 0·067	− 0·067	0·67
				$\chi^2 = 17·26$
				$g = 13$

hypothetical population would give χ^2 as great or greater than that observed.

When testing the correlations of Table 8.2 by finding the mean z', we implicitly assumed them to be samples from the same population and regarded the variations in r as due to random errors. It may be, however, that the variations are real racial differences, and that in general there is a correlation between cephalic index and

face form, which is sometimes positive and sometimes negative, so that the mean is nearly zero; we can now test this, using z' and χ^2. The values of $z'^2(N-3)$ are given in the fifth column of the table, and their sum gives a χ^2 of 17·26, which for 13 degrees of freedom lies between the 0·1 and 0·2 levels of significance. Thus the combined experience of Table 8.2 lends no support to the view that the two characters are associated, even after making allowance for the possibility of racial differences.

Having found a mean z' from a number of samples by the method of section 8.22, and found it to be significant, we may wish to

TABLE 8.3

CORRELATIONS BETWEEN HEAD LENGTH AND BREADTH

District	Number of Cases N	Correlation Coefficient	z'	$z'-\bar{z}'$	$(z'-\bar{z}')^2(N-3)$
Alexandria ..	643	+0·244	+0·249	−0·017 8	0·20
Cairo ..	802	+0·244	+0·249	−0·017 8	0·25
Canal ..	127	+0·330	+0·343	+0·076 2	0·72
Beheira ..	526	+0·213	+0·216	−0·050 8	1·35
Gharbiya ..	1 104	+0·316	+0·327	+0·060 2	3·99
Minufiya ..	717	+0·230	+0·234	−0·032 8	0·77

test if the individual values as a whole differ significantly among themselves, and the χ^2 distribution may be used for this. If z' is an individual transformed correlation, $\bar{z}'$ the mean and N the size of the individual sample,

$$\chi^2 = S(z'-\bar{z}')^2(N-3),$$

and the number of degrees of freedom (g) is one less than the number of samples (if the mean has been found from them). As an example, we will consider Table 8.3, which contains correlations between head length and breadth for Egyptians native to six districts; the data have been taken from a paper by Orensteen (1920). The weighted mean z' is + 0·266 8 (giving $r = +0·260$), and the sum of the last column is $\chi^2 = 7·28$; for 5 degrees of freedom, this lies almost exactly on the $P = 0·2$ level, and we must conclude that there is no evidence of any difference in correlation among the six districts

chosen. Comparing the Beheira and Gharbiya districts, we find the difference in z' is 0·111 with a standard error of

$$\pm\sqrt{\frac{1}{523}+\frac{1}{1\,101}}=\pm 0{\cdot}053;$$

this difference is about twice the standard error, and if the two samples came alone, it might be regarded as significant; but we must remember that it is just about the most significant one that could be taken from the six values, and bearing in mind the remarks in section 3.33 regarding significances of differences within groups of samples, we should compare the ratio 0·111/0·053, not with 2·0, the 0·05 level for a random pair, but with 2·8, the 0·05 level of the range for six samples (Table 3.3.)

Significances of Regression Constants

8.3. We have seen in section 8.1 how the analysis of variance technique and the z- and t-tests may be used for testing the significance of a correlation coefficient; here we shall extend these methods to testing a number of hypotheses concerning regression coefficients and lines. The hypotheses will be given in italics at the head of the sections.

8.31. *Hypothesis: That the population value of the regression coefficient is zero.*

This is the same as postulating a population with zero correlation, and the test of section 8.1 may be used. However, we shall express the equation in new terms.

In the notation of the last chapter, let the regression line of the sample be

$$(Y-\bar{y})=a(x-\bar{x}),$$

and the standard deviation of the residual deviations from the line be s, where

$$s^2={}_yv_r.$$

Then it is easy by means of equations (7.4) and (7.5) to transform equation (8.2) to

$$t=\frac{a}{\dfrac{s}{\sqrt{S(x-\bar{x})^2}}}$$

where s is estimated on $N-2$ degrees of freedom.

Since t has been described in section 5.21 as the ratio of a quantity to its standard error, we may say that the standard error of the regression coefficient in samples from a population with zero association is

$$\frac{s}{\sqrt{S(x-\bar{x})^2}},$$

and the method of section 5.2 may be used to test any significance.

8.32. *Hypothesis: That the population value of the regression coefficient is* α.

It is easy to show that the above expression for the standard error of the regression coefficient applies for any population value, α, so the test of the hypothesis is equivalent to using the t-test to see if the ratio of $a-\alpha$ to its standard error is significant for $N-2$ degrees of freedom.

In large samples where N can be substituted for $N-2$, the standard error of the regression coefficient reduces to

$$\sqrt{\frac{1-\rho^2}{N}} \cdot \frac{\alpha}{\rho},$$

where ρ is the population value of the correlation coefficient.

8.33. *Hypothesis: That a number of samples are from populations having the same regression coefficient. The residual variance of y is assumed to be constant.*

Consider the data of Table 8.4, taken from a paper by Glanville and Reid (1934). They are the crushing strengths of specimens of mortar and concrete made from seven different cements; specimens were broken at three different ages. These data form part of the result of an investigation into the use of tests on mortar as a guide to the behaviour of the cement in concrete. If the corresponding strengths of mortar and cement are plotted a marked correlation will be apparent, and we may ask if the slopes of the regression lines of concrete strength on mortar strength for the three ages are significantly different. Whether or not the means for the three ages are different does not here concern us, nor the possibility of any

variations in the residual variance of concrete strength from one age to another.

We may eliminate the mean strengths for the three ages by measuring the individual strengths as deviations from those means. Then we may either (1) fit one regression line to all the deviations, which is equivalent to fitting a series of parallel lines going through the means for the various ages, or (2) fit separate regression lines for the ages. The data are plotted in Fig. 11, where the parallel

TABLE 8.4

CRUSHING STRENGTH OF MORTAR AND CONCRETE, LB. PER SQ. IN. ÷ 10

Cement	One Day		Seven Days		Twenty-eight Days	
	Mortar x	Concrete y	Mortar x	Concrete y	Mortar x	Concrete y
A	263	138	750	507	895	679
B	493	278	936	653	1 066	818
C	137	49	453	425	632	651
D	477	293	893	662	1 100	842
E	233	104	545	437	716	603
F	568	350	797	735	897	832
G	230	141	631	439	846	724
$S_s(x - \bar{x}_s)^2$ ..	164 786		193 074		173 057	
$S_s(y - \bar{y}_s)^2$ ..	76 259		99 933		55 482	
$S_s(x - \bar{x}_s)(y - \bar{y}_s)$	111 205		117 648		82 646	

lines of system (1) are full lines and the separate ones of system (2) are dotted. The dotted lines fit the data more closely than the full ones in the sense that the sum of squares of deviations of concrete strength is less from the dotted lines. The separate regressions are only significantly different, however, if this difference in sums of squares is real after allowing for degrees of freedom absorbed in fitting and for random errors. This is tested by calculating residual variances. If the hypothesis is correct, the lines with separate slopes will give the same residual variance as those with the common slope, within the limits of sampling errors;* if the second residual variance is significantly less than the first, the hypothesis is rejected and it is inferred that the separate lines give a closer representation of the

* This follows from the proof by Fisher (1925) mentioned in section 7.23.

data than the common line and are significantly different in slope. We shall set out the process algebraically.

If there are u samples of $N_1\ N_2 \ldots$ individuals and we fit a separate regression line to each, viz.:

$$(Y - \bar{y}_1) = a_1(x - \bar{x}_1),$$
$$(Y - \bar{y}_2) = a_2(x - \bar{x}_2),$$
$$\ldots\ldots\ldots\ldots$$

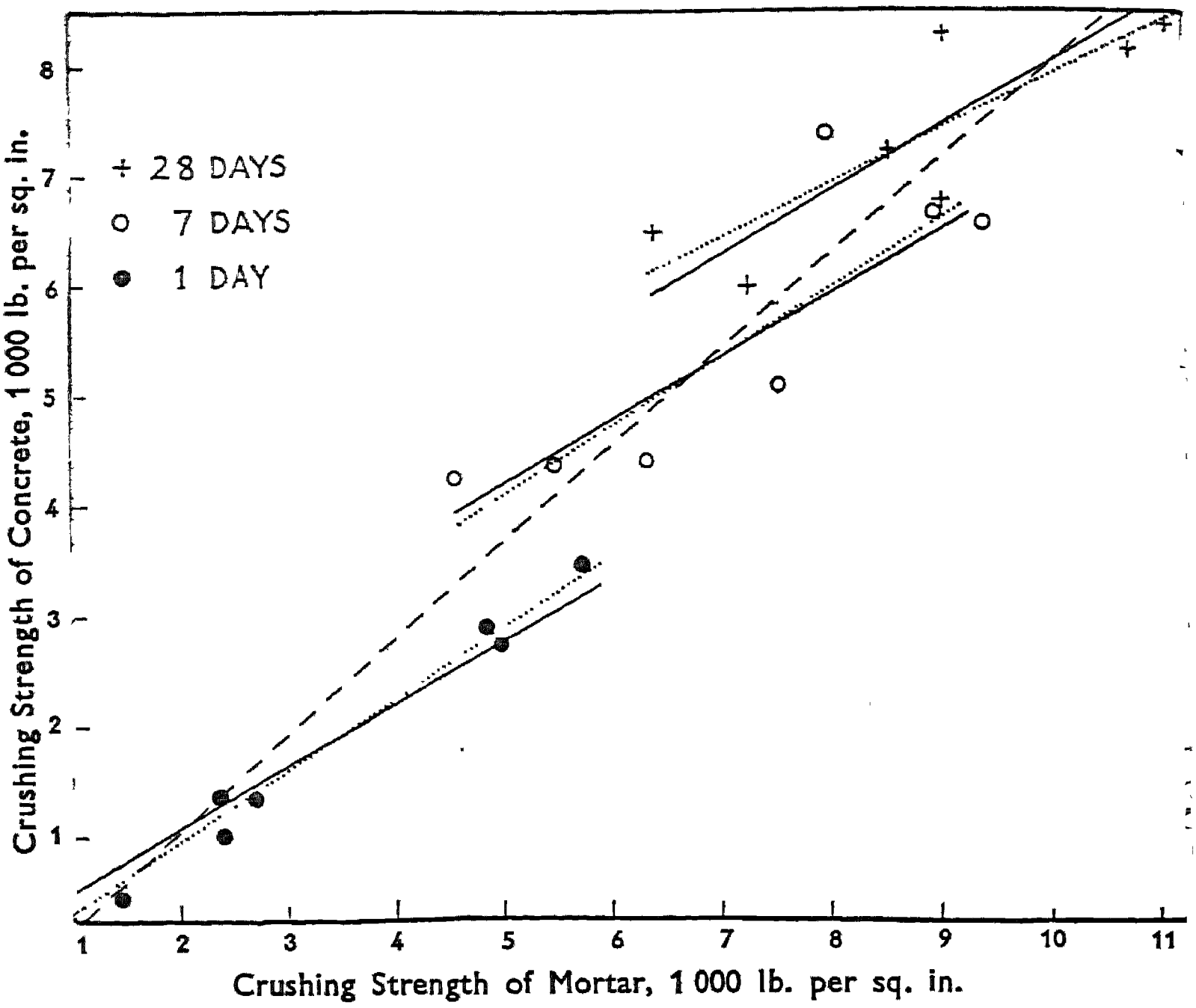


Fig. 11

the residual sums of squares and corresponding degrees of freedom after fitting the lines are given by equation (7.92) as:

$$\textit{Sums of squares:}\quad S_1(y - \bar{y}_1)^2 - \frac{[S_1(x - \bar{x}_1)\,(y - \bar{y}_1)]^2}{S_1(x - \bar{x}_1)^2},$$
$$S_2(y - \bar{y}_2)^2 - \frac{[S_2(x - \bar{x}_2)\,(y - \bar{y}_2)]^2}{S_2(x - \bar{x}_2)^2} \quad \ldots \quad (8.4)$$
$$\text{etc.,} \qquad \text{etc.,}$$

$$\textit{Degrees of freedom:}\quad N_1 - 2, \quad N_2 - 2. \quad \text{etc.,}$$

where $S_1\ S_2\ \ldots$ are summations over all individuals in the respective samples. The sum of these sums of squares divided by $(N_1 + N_2 + \ldots - 2u)$, the sum of the degrees of freedom, provides an estimate of the residual variance of y in the population, which is assumed to be the same for all samples.

Now, assuming the samples to have one regression coefficient but different means of y for a given value of x, the residual sum of squares from the regression lines with common slope, and the degrees of freedom are:

Sum of squares:

$$S_1(y-\bar{y}_1)^2 + S_2(y-\bar{y}_2)^2 + \ldots - \frac{[S_1(x-\bar{x}_1)(y-\bar{y}_1) + S_2(x-\bar{x}_2)(y-\bar{y}_2) + \ldots]^2}{S_1(x-\bar{x}_1)^2 + S_2(x-\bar{x}_2)^2 + \ldots} \quad (8.41)$$

Degrees of freedom: $N_1 + N_2 + \ldots - u - 1$.

The degrees of freedom are made up of the $N - 1$ contributed by each sample minus the one absorbed in fitting the common regression. The sum of squares divided by the degrees provides another estimate of the residual variance of y.

We cannot test the difference between the two variances obtained from (8.4) and (8.41) for significance by the z test, because the variances are not independent. However, if we subtract the sum of squares of (8.4) from that of (8.41) and divide by the corresponding difference in degrees, we have an estimate of the variance that is independent of the estimate of (8.4), and may be compared with it. The differences in sum of squares and degrees of freedom are:

Sum of squares:

$$\frac{[S_1(x-\bar{x}_1)(y-\bar{y}_1)]^2}{S_1(x-\bar{x}_1)^2} + \frac{[S_2(x-\bar{x}_2)(y-\bar{y}_2)]^2}{S_2(x-\bar{x}_2)^2} + \ldots - \frac{[S_1(x-\bar{x}_1)(y-\bar{y}_1) + S_2(x-\bar{x}_2)(y-\bar{y}_2) + \ldots]^2}{S_1(x-\bar{x}_1)^2 + S_2(x-\bar{x}_2)^2 + \ldots}, \quad (8.42)$$

Degrees of freedom: $u - 1$.

Then z may be calculated from the ratio of the variances estimated from (8.42) and (8.4) and tested for significance for degrees of freedom, $n_1 = u - 1$, $n_2 = N_1 + N_2 + \ldots - 2u$.

For the data of Table 8.4, the total sum of squares of concrete strengths from equation (8.4) is*

$$76\,259 - \frac{(111\,205)^2}{164\,786} + \ldots = 45\,471,$$

and there are 15 degrees of freedom, giving a variance of 3 031. The sum of squares in equation (8.42) is

$$\frac{(111\,205)^2}{164\,786} + \ldots - \frac{(311\,499)^2}{530\,917} = 3\,441,$$

and for 2 degrees of freedom this gives as an estimate of variance the value 1 720. This is not greater than the variance from (8.4), and the slopes of the lines are not significantly different.

8.34. *Two samples are from populations having the same regression coefficient.*

The above test may be reduced to the t-test when there are only two samples. The variance from (8.42) is, in this instance, based on one degree of freedom, the square root of its ratio to the variance from (8.4) is t.

If the residual variance estimated from (8.4) is written s^2 and the regression coefficients of the two samples are a_1 and a_2, it is easy to show from equations (7.5), (8.4) and (8.42) that

$$t = \frac{a_1 - a_2}{s\sqrt{\dfrac{1}{S_1(x - \bar{x}_1)^2} + \dfrac{1}{S_2(x - \bar{x}_2)^2}}},$$

where s is estimated from $N_1 + N_2 - 4$ degrees of freedom.

This test is parallel to that given in section 5.3 for the difference between two means. It follows from this that the standard error of the difference between the two regression coefficients may be derived from those of the separate coefficients by the standard method, using equation (3.1), p. 72.

As an example we may consider the seven and twenty-eight day results of Table 8.4. The residual variance s^2 is obtained from equation (8.4) applied to the last two ages, and is 4 426 based on 10 degrees of freedom. The standard error of the difference between the two regression coefficients is

* All results for this example are in units of 10 lb. per sq. in.

$$\sqrt{4\,426\left(\frac{1}{193\,074}+\frac{1}{173\,057}\right)}=0{\cdot}220.$$

The regression coefficients are

$$\frac{117\,648}{193\,074}=0{\cdot}609\,3 \quad \text{and} \quad \frac{82\,646}{173\,057}=0{\cdot}477\,6.$$

Hence, $t=0{\cdot}6$, and for 10 degrees of freedom this is near the 0·6 level; the difference in regression coefficients is not significant.

8.35. *Hypothesis: That a number of samples are from populations having the same regression line. The residual variance of y is assumed to be constant.*

We are postulating here, not only that the true regression lines for the separate samples are parallel, but that they *coincide*. The true means and variances of the independent variable x need not be equal for the populations, so those of y are not necessarily equal, but if our hypothesis is true, the true mean value of y and the variance for any given value of x must be the same for all samples. This hypothesis is not quite the same as postulating that the samples come from the same population. Again, we are not considering the possibility of the residual variance of y changing from sample to sample, and so are not testing it.

The general procedure is the same as that used in section 8.33. We fit the separate lines shown dotted in Fig. 11 and estimate a residual sum of squares and variance from equation (8.4). The common line is shown in Fig. 11 by a broken line; it is fitted to the deviations of the values from the grand means for all samples. If $\bar{x}$ and $\bar{y}$ are the grand means and S is the summation over all individuals, we have for the residuals from this line:

$$\left.\begin{aligned} &\textit{Sum of squares:}\quad S(y-\bar{y})^2-\frac{[S(x-\bar{x})\,(y-\bar{y})]^2}{S(x-\bar{x})^2},\\ &\textit{Degrees of freedom:}\quad N_1+N_2+\ldots-2. \end{aligned}\right\} \quad (8.43)$$

The difference between the sums of squares in (8.43) and (8.4) divided by the difference in degrees of freedom, $2u-2$, is the independent estimate of variance that may be compared with the residual variance from (8.4). There is no further algebraic reduction of the expressions.

We shall test the hypothesis on the data of Table 8.4 analysing

the variance of the concrete strengths. The sum of squares of deviations from the common line is 111 792, the sum of squares of deviations from the separate lines is 45 471 and the difference is 66 321 for $19 - 15 = 4$ degrees, giving as an estimate of variance the value 16 580. The residual variance based on 15 degrees is 3 031, and since this is significantly less than the estimate based on four degrees ($z = 0{\cdot}85$ lies beyond the $0{\cdot}01$ point), we conclude that the regression lines differ. In this instance, since we have shown in section 8.33 that the lines do not differ significantly in slope, we may say that they differ only in level.

To test for a difference in level, assuming the slope to be constant, it would be legitimate to compare the residual after fitting a number of lines having different means but the same slope, with the residual after fitting a common line. Such a test is used in section 11.11.

Effect of Non-uniformity of Variance

8.36. The tests for these hypotheses are based on the assumption that the residual variances are the same for the populations sampled. It is probable that the tests are appreciably affected if the residual variances change from one of the populations to another by more than a small amount. Where heterogeneity of variance is suspected, and a result of importance is very near the level of significance, it is well to test this assumption by the methods of sections 5.4 and 5.5.

For example, the separate residual variances for concrete of the three ages in Table 8.4 are 243, 5 649 and 3 203, each being based on 5 degrees, and the first is significantly less than the other two. In consequence, some little doubt may be thrown on the inferences of the sections 8.33 and 8.35. We do not propose attempting to extend the statistical analysis to deal with such instances, and if the matter is of importance suggest either that more data should be procured or that here the hypotheses could be tested on the 7- and 28-day values only.

SHEPPARD'S CORRECTIONS

8.4. In calculating correlation coefficients from samples of grouped data, Sheppard's corrections are often used in determining σ_x and σ_y, and they tend to increase the value of r. For this reason it is better to ignore them in making tests of significance, particularly when the grouping is broad, although their use probably leads to an estimate of r which is nearer the population value.

CHAPTER IX

NON-LINEAR REGRESSION AND THE MEASUREMENT OF ASSOCIATION

The Correlation Ratio

9.1. We have seen that the correlation coefficient measures the proportion of total variance of a quantity associated with a straight regression line, but if the true regression is not linear, i.e. if the array means of one quality tend to lie on a curve which deviates

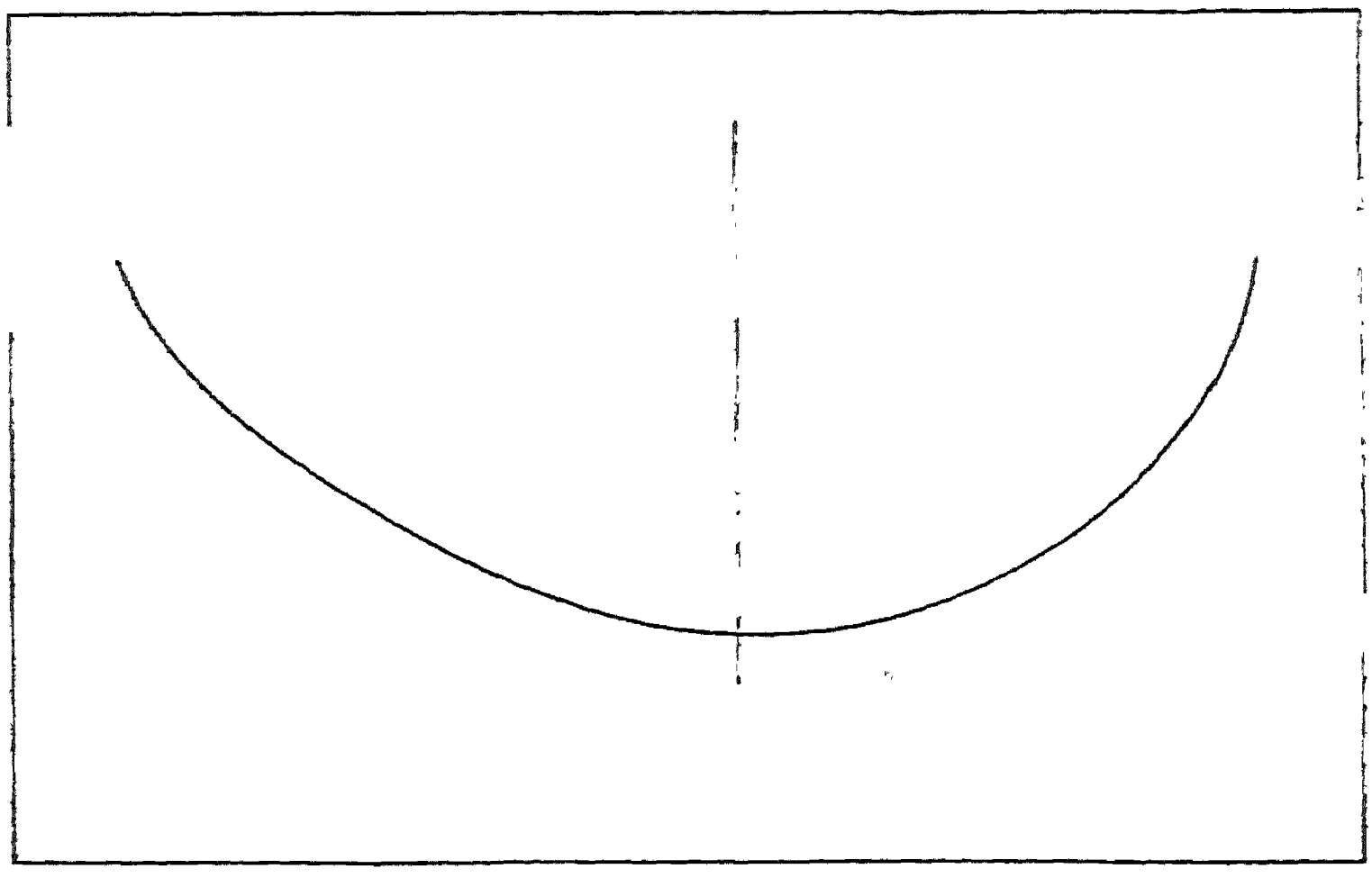


Fig. 12.

sensibly from a straight line, the correlation coefficient may be a very inadequate representation of the true state of affairs. It is possible to imagine a case in which the two regression curves are somewhat like those shown in Fig. 12, one being a straight line and the other a curve more or less symmetrical about a vertical axis. The correlation coefficient would be nearly zero (for the best straight line that could be drawn to fit the curve would be parallel to the x-axis), although there might be a high association as expressed by the curve; we are not aware of any such extreme case, but the same considerations apply when the curvature is less marked. In such instances, the only thing to do is to present the correlation table, to give the array

means and perhaps to fit a curve to them; methods of performing this last operation will be discussed later. Professor K. Pearson (1905) has proposed a constant called the *correlation ratio* as a measure of association analogous to the correlation coefficient, and suitable for use when the regression is not linear. If x is an individual observation of one variate, $\bar{x}$ the grand mean and $\bar{x}_s$ the mean of any individual array, the *correlation ratio of x on y* is η_{xy}, where

$$1 - \eta_{x_y}^2 = \frac{S(x - \bar{x}_s)^2}{S(x - \bar{x})^2}$$

and the *correlation ratio of y on x* is η_{y_x}, where . . . . (9.1)

$$1 - \eta_{y_x}^2 = \frac{S(y - \bar{y}_s)^2}{S(y - \bar{y})^2},$$

S being the summation over all individuals. If we now refer back to Table 6.3, p. 130, we see that

$$(1 - \eta_{x_y}^2) = \frac{\textit{Sum of squares of deviations from array means of } x}{\textit{Total sum of squares}}$$

and thus η is analogous to the correlation coefficient r, since from Table 7.61, p. 158, we see that

$$(1 - r^2) = \frac{\textit{Sum of squares of deviations from linear regression}}{\textit{Total sum of squares}}.$$

Hence r^2 is the proportion of the total sum of squares of x associated with y when a straight line is fitted, and η^2 is the proportion associated when a series of array means or curve is fitted; just as there are two regression lines (for which r happens to be the same), there are two series of array means which require separate correlation ratios. If the sample is large compared with the number of arrays and we may neglect the degrees of freedom absorbed, and if σ_x is the standard deviation of x, and σ'_x that of the residual deviations, we have as an approximation,

$$\text{for linear regression} \quad \sigma_x'^2 = (1 - r^2)\sigma_x^2,$$
$$\text{for curved regression} \quad \sigma_x'^2 = (1 - \eta_{x_y}^2)\sigma_x^2.$$

In large samples, therefore, η may be interpreted as a measure of association on practically the same scale as r. We have, however, made the *proviso* that the number of arrays shall be small compared

with the size of the sample (say one-fiftieth). The usefulness of this ratio is increased by the fact that the quantity y does not enter into the expression for η_{x_y}, so that the second variate need not even be quantitative; we have thus a measure of association which may be used in such analyses as those of Tables 6.1 and 6.5, pp. 127 and 133. From those tables we see that the correlation ratios are:
Number of ovules per ovary,

$$1 - \eta^2 = \frac{3\,026{\cdot}350}{5\,379{\cdot}775}, \quad \eta^2 = 0{\cdot}437, \quad \eta = 0{\cdot}66,$$

Length of cuckoos' eggs,

$$1 - \eta^2 = \frac{2\,356{\cdot}21}{3\,429{\cdot}70}, \quad \eta^2 = 0{\cdot}313, \quad \eta = 0{\cdot}56,$$

and the strength of association in the former case comes somewhere between those of the pistils and stamens of the late and early flowers (Tables 7.2 and 7.3, pp. 145 and 146), while in the second case it is practically the same as the relationship between the pistils and stamens for early flowers.

One disadvantage of the correlation ratio, however, is that it is not independent of the number of groups from which it is calculated —that it is not entirely a property of the population nor even of the sample, but depends on its arrangement into arrays. If there are m arrays each with n observations (so that the total in the sample is $N = mn$), if ${\sigma_s}^2$ and ${\sigma_r}^2$ are the squares of the standard deviation of array means and residual deviations in the infinite population, and if v_s and v_r are the corresponding variances, we see from Table 6.3, p. 130, that

$$\eta^2 = \frac{(m-1)v_s}{m(n-1)v_r + (m-1)v_s} \rightarrow \frac{n{\sigma_s}^2 + {\sigma_r}^2}{n{\sigma_s}^2 + {\sigma_r}^2 + \dfrac{N-m}{m-1}{\sigma_r}^2} \quad . \quad (9.2)$$

and this shows that η^2 depends on n and m. If the number in each array (n) is so large that σ_r^2 may be neglected in comparison with $n\sigma_s^2$, and n may be written for $(n-1)$, equation (9.2) may be written

$$\eta^2 \rightarrow \frac{1}{1 + \dfrac{m}{m-1} \cdot \dfrac{{\sigma_r}^2}{{\sigma_s}^2}} \quad . \quad . \quad . \quad . \quad (9.21)$$

and the larger m is, the smaller is $m/(m-1)$, or the larger is η^2. Thus it is important to keep the number of groups small so as to avoid over-emphasising the association.

Since η^2 is related to a ratio of variances, it may be used instead of z as a test for significance of association. However, such a test is not convenient and is seldom used, and we shall not deal with it further.

Tests for Non-Linearity of Regression

9.2. We shall consider the data of Table 9.1 (Zinn, 1923), relating the loaf volume and percentage protein in a number of Ohio commercial wheats. The interest lies in predicting the loaf volume from the protein content, and so we consider the regression of the former on the latter. The correlation coefficient (found without Sheppard's corrections) is $+0{\cdot}529\,78$, showing that the loaf volume does, on the whole, increase with the protein content. If the table is examined, however, there seems to be a tendency for the loaf volume to remain practically constant for wheats of low protein content, and for the relationship between the two quantities to be more marked for wheats with more protein, i.e. for the regression to be non-linear. Is this apparent tendency real, or is it due to errors of random sampling? This section is concerned with tests of significance of such departures from linearity.

The general method is based on the analysis of variance. If the regression is non-linear, it will be almost completely expressed by the mean values of loaf volume for the separate arrays, and the residual variance calculated by the methods of section 6.4 will be as low as is possible. A straight regression line fitted to the same data will have a significantly greater residual variance, because the residual will contain variance due to the deviations from linearity. On the other hand, if the regression is linear, it may easily be shown that these two residual variances will be estimates of the same variance. The test for non-linearity therefore resolves itself into a test of significance of the difference between the two estimates. However, they are not independent, and so cannot be compared directly by the z-test, but the difference between the sums of squares divided by the difference in degrees of freedom is independent of the residual variance of deviations from the array means, and these two estimates of variance may be compared.

The sums of squares and degrees of freedom for a sample of

TABLE 9.1.—PROTEIN CONTENT OF WHEAT AND LOAF VOLUME

Loaf Volume, c.c.	Protein Content, per cent. 9·0	10·0	11·0	12·0	13·0	14·0	15·0	16·0	17·0	Totals
1 600–	2	—	1	—	1	—	—	—	—	4
1 700–	—	2	1	—	2	—	—	—	—	5
1 800–	6	5	5	2	1	—	—	—	—	19
1 900–	5	4	6	7	—	—	—	—	—	22
2 000–	2	3	9	10	2	1	—	—	—	27
2 100–	1	—	3	3	5	2	1	—	—	15
2 200–	—	—	1	2	—	2	1	—	—	6
2 300–	—	—	—	—	—	—	—	1	1	2
Totals ..	16	14	26	24	11	5	2	1	1	100

TABLE 9.2

Source of Variation	Sums of Squares	Degrees of Freedom
Linear regression	$r^2 \underset{s}{S} S S'(y - \bar{y})^2$	1
Deviations from linear regression	*difference*	$m - 2$
Between arrays	$\underset{s}{S} n_s(\bar{y}_s - \bar{y})^2 = \eta^2 S(y - \bar{y})^2$	$m - 1$
Residual within arrays	$S S'(y - \bar{y}_s)^2 = (1 - \eta^2) S(y - \bar{y})^2$	$N - m$
Total	$\underset{s}{S} S S'(y - \bar{y})^2$	$N - 1$

N observations divided into m arrays, with n_s observations in the sth array is given in Table 9.2, where all the symbols have the meanings given to them in Chapters VI and VII. All the sums of squares given directly are taken from Tables 6.3 (writing y for x and adapting it for non-uniform array totals) and 7.6. It is recommended that the sum of squares in the second row should be obtained by subtraction.

TABLE 9.3

ANALYSIS OF VARIANCE OF LOAF VOLUMES

(Units, 10 000 c.c.^{2})

Source of Variation		Sums of Squares		Degrees of Freedom		Mean Variance
Linear regression	Between arrays	64·66	78·21	1	8	64·66
Deviations from linear regression		13·55		7		1·94
Residual within arrays ..		152·15		91		1·67
Total		230·36		99		—

The analysis for the data of Table 9.1 is in Table 9.3. The variances to be compared give a value of z of

$$\tfrac{1}{2}\log_e\left(\frac{1{\cdot}94}{1{\cdot}67}\right) = 0{\cdot}075,$$

and since this is well below the 5 per cent. level of significance, there is no evidence that the regression departs from linearity.

It will be noticed that the distribution of protein content is far from normal, but that does not invalidate our test, which depends only on the assumption that the distribution of loaf volume within the arrays is practically normal. As far as can be judged from a sample of this size, this assumption is justified.

POLYNOMIAL REGRESSION CURVES

9.3. Sometimes it is convenient to represent the regression of y on x by a smooth curve described by an equation with several constants

that have to be determined from the data. From the statement in the footnote in section 7.23 it will be appreciated that such an equation having only a few constants may absorb fewer degrees of freedom than a number of array means, leaving more degrees for the estimation of the residual. When there are only a few observations, they cannot be grouped in arrays and the use of an equation may be essential in the analysis of variance.

The most useful form of equation is the polynomial

$$Y = a + bx + cx^2 + \ldots \quad . \quad . \quad . \quad . \quad (9.3)$$

and according to the method of least squares the constants a, b, c, etc., may be obtained by solving a series of linear equations

$$\left.\begin{aligned} Sy &= aS_1 + bSx + cSx^2 + \ldots \\ Sxy &= aSx + bSx^2 + cSx^3 + \ldots \\ Sx^2y &= aSx^2 + bSx^3 + cSx^4 + \ldots \end{aligned}\right\} \quad . \quad . \quad (9.4)$$

using as many equations as there are constants. The summations are taken over all individuals and may be determined by fairly straightforward extensions of the methods of sections 1.3 and 7.5. S_1 is N, Sx and Sy are N times the respective means and the other sums are N times various moments and product moments measured about the origin. Any convenient arbitrary origin may be used in finding the summations provided the same origin is used in expressing the result in (9.3). If the data are grouped into arrays, the array means $\bar{y}_s$ may be used instead of the individual values, and the summations are then $\sum_s n_s\bar{y}_s$, $\sum_s n_s$, $\sum_s n_s x_s$, etc., using the same notation as before. A similar series of equations obtained by interchanging x and y in equations (9.3) and (9.4) gives the regression of x on y.

The polynomial equation is very adaptable, and according to the values of the constants a, b, c, etc., it may take a wide variety of forms with only a comparatively few constants. Indeed, for practical purposes it is a suitable way of expressing nearly all kinds of regression except those in which the curve of y undulates and has more than two or three maxima and minima.

In the analysis of variance given in section 9.2, a polynomial equation may be used instead of the array means, and it absorbs as many degrees of freedom as there are constants, the constant a accounting for the one degree previously attributed to the grand

mean of y. Curves of the first, second, third and higher orders may be fitted successively if desired, and as more and more terms are used, the equation fits the observations more and more closely, until ultimately an equation with as many constants as there are observations does so perfectly. This progressive improvement in the closeness of fit may be seen by finding the deviations of the observed values from those given by the regression and squaring and adding them, i.e. by finding $S(y - Y)^2$. As more and more terms are used this sum of squares decreases until when the fit becomes perfect, it becomes zero. If the constants are determined by the method of least squares and the fluctuations in y are purely random, the reduction in sum of squares at each stage tends to be the same; it is the variance associated with one degree of freedom, differing from the true variance only by the extent of the random errors. If the data show a comparatively simple form of variation, predominating over the random fluctuations, the earlier terms of the equation, which express the simpler movements, reduce the sum of squares by an amount significantly greater than the later ones. When sufficient terms are included to express this slow movement of y with x, and the residual deviations are random, each of the remaining possible terms reduces the remaining sum of squares by the same amount, within the limits of random errors. Thus, at each stage in fitting equations with more terms, the reduction in residual sum of squares due to one degree of freedom can be compared with the variance estimated from the residual; if the former is significantly greater than the latter, the last term expresses a real feature of the trend. When the stage is reached that the variances associated with the last one or two terms are not significantly greater than that estimated from the residual, it is presumed that the regression has been adequately expressed and the deviations are random. This presumption is only justified, however, if the polynomial is the form appropriate for describing the regression. This technique of fitting and testing curves should not be followed blindly, and a certain amount of judgment and reference to a diagram may be necessary; it is possible that the first term or two may account for very little of the variance, but that some of the later ones may be very important; this would be the case in the curve of Fig. 12, where the linear term would be very small but the second one would be important. On the other hand, if a very large number of terms is required, it may be because the polynomial form is not suitable. A regression equation fitted to

any data must not only satisfy these statistical tests, but it must appear appropriate when plotted on a diagram.

Every time an equation with an additional constant is tried, equations (9.4) have to be solved afresh and new values of all the constants obtained. Thus, a third order curve requires different values of a and b from those appropriate to the second order. The equation (9.3) may be expressed in a so-called *orthogonal* form consisting of a number of terms of successively higher orders in x, each term being multiplied by a constant to be determined from the data. These terms are independent in that they may be determined one at a time, and the addition of the higher orders does not alter the constants of the lower orders previously determined. This course is easiest when the values of the independent variable are at equal intervals and there is one value of the dependent for each value of the independent variable, as in a time series. Fisher (1936) describes a convenient system for doing this, and we shall illustrate it by fitting a curve of the third degree to the protein content data of Table 8.1.

9.31. We shall call the mean protein content (the observed values) y, and the year measured from the middle year 1910, t (i.e. for 1908, $t = -2$), and will fit a curve of the form,

$$Y = a + bt + ct^2 + dt^3 \quad . \quad . \quad . \quad . \quad (9.5)$$

to the 29 values ($N = 29$).*

This Fisher transforms to

$$\left.\begin{aligned} Y &= A + BT_1 + CT_2 + DT_3, \\ \text{where} \qquad & \\ T_1 &= (t - \bar{t}) = t \text{ (since } \bar{t} = 0), \\ T_2 &= t^2 - \frac{N^2 - 1}{12} = t^2 - 70, \\ \text{and} \qquad & \\ T_3 &= t^3 - \frac{3N^2 - 7}{20}t = t^3 - 125{\cdot}8t. \end{aligned}\right\} \quad . \quad . \quad (9.6)$$

* If there is an even number of years, $t = 0$ must still be at the centre, so that the values of t must be $+ 0{\cdot}5, + 1{\cdot}5, \ldots - 0{\cdot}5, - 1{\cdot}5 \ldots$

The constants are given by the relations

$$A = \frac{1}{N}Sy = \bar{y},$$

$$B = \frac{12}{N(N^2-1)}SyT_1 = \frac{1}{2\,030}Syt,$$

$$C = \frac{180}{N(N^2-1)(N^2-4)}SyT_2 = \frac{1}{113\,274}\{Syt^2 - 70Sy\}, \quad (9.7)$$

$$D = \frac{2\,800}{N(N^2-1)(N^2-4)(N^2-9)}SyT_3$$

$$= \frac{5}{30\,292\,704}\{Syt^3 - 125{\cdot}8Syt\}.$$

The summations Sy, Syt, etc., have to be found from the data, and Fisher gives a convenient method of evaluating them by a series of additions, but for such a small number of observations as we have, it is not very much trouble to obtain the summations directly, particularly if a calculating machine is available. For the protein percentages they are,

$Sy_1 = 411{\cdot}48$	$Sy_2 = 240{\cdot}78$
$Sy_1t = +\ 342{\cdot}98$	$Sy_2t = -\ 195{\cdot}36$
$Sy_1t^2 = 28\,408{\cdot}36$	$Sy_2t^2 = 17\,674{\cdot}16$
$Sy_1t^3 = +\ 49\,329{\cdot}62$	$Sy_2t^3 = -\ 26\,716{\cdot}68.$

If ${}_1y_1, {}_2y_1, {}_3y_1 \ldots {}_{29}y_1$ are the individual readings,

$$Sy_1t = ({}_{29}y_1 - {}_1y_1)\cdot 14 + ({}_{28}y_1 - {}_2y_1)\cdot 13 + \ldots + ({}_{16}y_1 - {}_{14}y_1)\cdot 1,$$
$$Sy_1t^2 = ({}_{29}y_1 + {}_1y_1)\cdot 14^2 + ({}_{28}y_1 + {}_2y_1)\cdot 13^2 + \ldots + ({}_{16}y_1 + {}_{14}y_1)\cdot 1^2,$$
$$Sy_1t^3 = ({}_{29}y_1 - {}_1y_1)\cdot 14^3 + ({}_{28}y_1 - {}_2y_1)\cdot 13^3 + \ldots + ({}_{16}y_1 - {}_{14}y_1)\cdot 1^3,$$
etc. etc.

when the values of t are symmetrically placed about the zero; the labour is diminished if the terms of the brackets are found first, giving two series of fourteen values, one for use when multiplying by the odd powers of t, and the other when multiplying by the even powers.

Substituting the above values in (9.7) we obtain,

$$A_1 = \frac{411{\cdot}48}{29} = 14{\cdot}188\,97, \qquad A_2 = \frac{240{\cdot}78}{29} = 8{\cdot}302\,76,$$

$$B_1 = \frac{342{\cdot}98}{2\,030} = 0{\cdot}168\,96, \qquad B_2 = \frac{-195{\cdot}36}{2\,030} = -0{\cdot}096\,24$$

$$C_1 = \frac{-395{\cdot}24}{113\,274} = -0{\cdot}003\,489\,2, \quad C_2 = \frac{819{\cdot}56}{113\,274} = 0{\cdot}007\,235\,2,$$

$$D_1 = \frac{5 \times 6\,182{\cdot}736}{30\,292\,704} = 0{\cdot}001\,020\,5,$$

$$D_2 = \frac{-5 \times 2\,140{\cdot}392}{30\,292\,704} = -0{\cdot}000\,353\,3.$$

These constants can be substituted in equation (9.6).

It is not necessary to calculate the residual deviations explicitly, for the succesive terms reduce the sums of squares by the following amounts:

$$NA^2, \quad \frac{N(N^2-1)}{12}B^2, \quad \frac{N(N^2-1)(N^2-4)}{180}C^2,$$

$$\frac{N(N^2-1)(N^2-4)(N^2-9)}{2\,800}D^2, \text{ etc.}$$

The first of these has been encountered before, for it is the ordinary way of correcting a sum of squares of deviations from some origin to a sum of squares of deviations from the mean;

$$NA^2 = N\bar{y}^2 \quad \text{and} \quad S(y-\bar{y})^2 = Sy^2 - N\bar{y}^2,$$

and the other terms may be regarded as correcting to a moving mean, the first to a mean which gradually increases (or diminishes) according to a straight line law, for

$$\frac{N(N^2-1)}{12}B^2 = r^2S(y-\bar{y})^2,$$

and the others to a mean the movements of which become more complicated as more and higher terms in the regression are used.

We are only concerned with the analysis of sums of squares of deviations from the means, and these are given in the "Total" row of Table 9.4 for the two series of protein contents, while the variances of the terms in the regression are given in higher rows. The variances of the second and third terms are both greater than the residuals, and we may test their significance by finding

$$z = \tfrac{1}{2}\log_e\left(\frac{1{\cdot}38}{0{\cdot}494}\right), \text{ etc.,}$$

and seeing if they are above the 0·05 level when $n_1 = 1$ and $n_2 = 25$, or alternatively by using the result of section 6.5 and finding

$$t = \sqrt{\frac{1 \cdot 38}{0 \cdot 494}}, \text{ etc.,}$$

and testing their significance for 25 degrees of freedom. Fisher's tables of t give 2·060 lying on the 0·05 level of significance, so that the first order terms are both real (we arrived at this conclusion in

TABLE 9.4

ANALYSIS OF VARIANCE

Source of Variation	Degrees of Freedom	Series 1		Series 2	
		Sum of Squares	Variance	Sum of Squares	Variance
First order regression	1	57·95	57·95	18·80	18·80
Second order regression ..	1	1·38	1·38	5·93	5·93
Third order regression ..	1	6·31	6·31	0·76	0·76
Residual	25	12·34	0·494	12·06	0·482
Total ..	28	77·98	—	37·55	—

section 8.1 when finding the correlation coefficient), and so are the third order terms of the first series and the second order terms of the second.

To test whether the cubic regression line as a whole gives a better fit than a linear one, we find

$$z = \tfrac{1}{2} \log_e \frac{3 \cdot 84}{0 \cdot 494} = 1 \cdot 025$$

for the first series and

$$z = \tfrac{1}{2} \log_e \frac{3 \cdot 34}{0 \cdot 482} = 0 \cdot 968$$

for the second series, and the degrees of freedom are $n_1 = 2, n_2 = 25$. Fisher gives for such a sample the 1 per cent. point of z as 0·858 5,

so the cubic regression is significantly different from the linear one in both instances.

It may be argued that there is no reason why we should stop at the third order term, but Fig. 13 shows that the data are quite well followed by the cubic curves, and from Table 9.4 we see that the residual variances for the two series are practically equal,

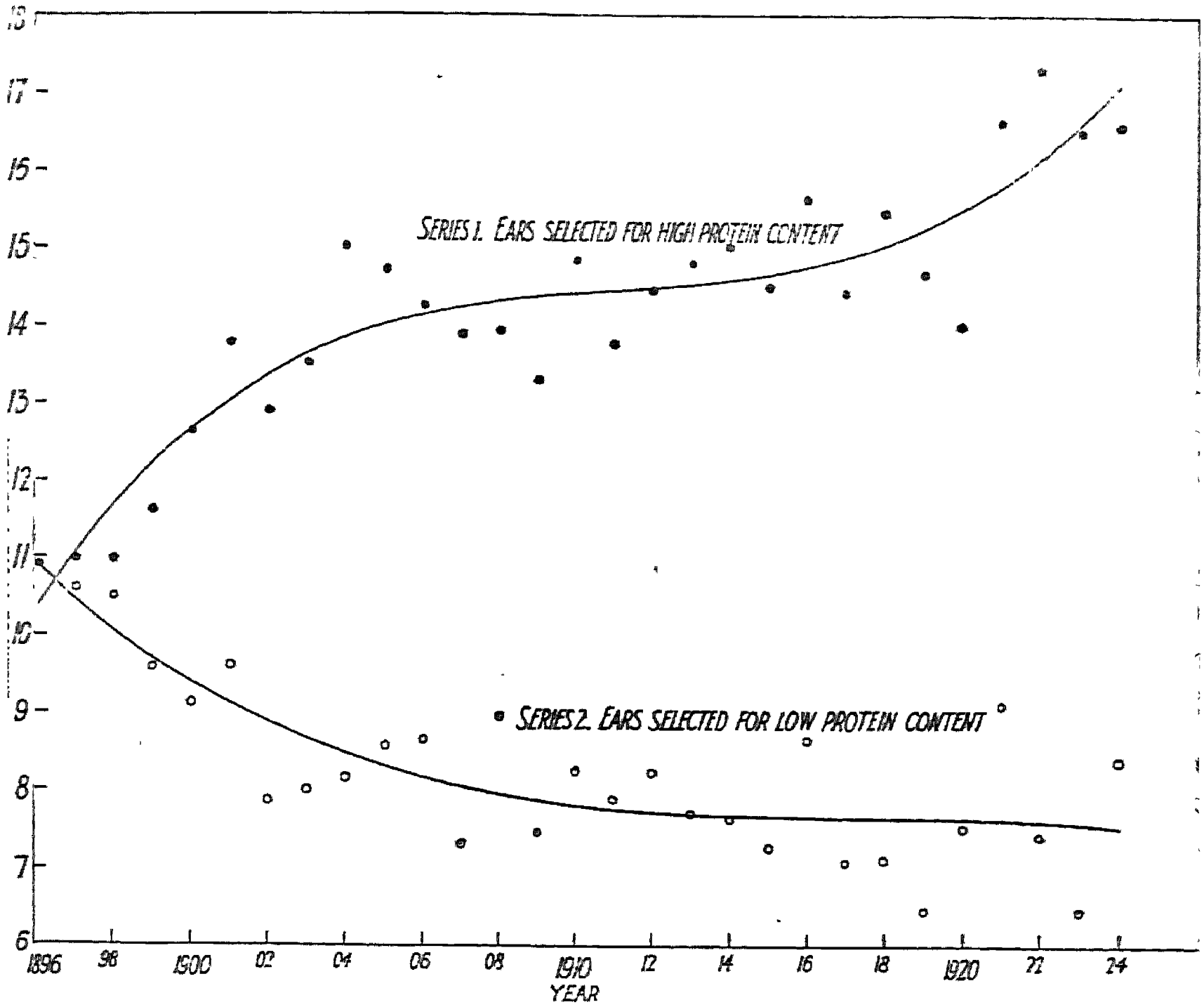


FIG. 13.

suggesting that they may both be a result of the same random causes. We shall use these data later, and assume that cubic equations eliminate sufficiently the systematic trend in protein content.

9.32. There are forms of equation other than the polynomial that may be appropriate in some instances; for example, equations involving trigonometric functions are suitable for expressing many types of periodic fluctuation with time. For smoothing economic time series, many special methods have been devised, and are often

favoured either because they describe complicated movements more readily than a simple polynomial does or because they are a mathematical expression of some economic generalisation. We cannot deal with these here, but would remark that many of these curves cannot be fitted by least squares or some equivalent method, so it is not possible to say how many degrees of freedom are absorbed, and the criteria of fit developed in this chapter may not be applied. In the absence of alternative criteria, this may be a disadvantage, although on the other hand it may perhaps be argued that in the absence of criteria for testing whether a suitable *form* of equation has been chosen, it is idle to place too much emphasis on criteria for testing whether the number of terms in the equation is appropriate.

9.4. It may be well to emphasise the fact that since all the tests of the preceding sections are extensions of the analysis of variance, they are based on the assumptions, discussed in section 6.7, of normality and homogeneity of the residual variations.

Contingency

9.5. In section 4.4 we have described the formation of contingency tables when the two characters can only be expressed in broad qualitative groups, and how the existence of association may be tested by calculating χ^2, but we have not given any method of expressing the strength of association. The best constant is probably Pearson's (1904) mean square contingency coefficient.

Consider the sth row and tth column in a contingency table, let the number of individuals in the cell common to both be n_{st}, the total in the sth row be $n_{s.}$, that in the tth column be $n_{.t}$, and the grand total be N. Then the expected frequency in the cell is $n_{s.}n_{.t}/N$, and we have seen that

$$\chi^2 = \sum_s \frac{\left(n_{st} - \dfrac{n_{s.}n_{.t}}{N}\right)^2}{\dfrac{n_{s.}n_{.t}}{N}}.$$

The *mean square contingency* is defined as

$$\phi^2 = \frac{\chi^2}{N} = \sum_s \frac{\left(n_{st} - \dfrac{n_{s.}n_{.t}}{N}\right)^2}{n_{s.}n_{.t}} \quad . \quad . \quad . \quad (9.8)$$

and the *mean square contingency coefficient* as

$$C_1 = \sqrt{\frac{\phi^2}{1 + \phi^2}} \quad . \quad . \quad . \quad . \quad . \quad (9.81)$$

This latter constant was chosen because as the grouping becomes very fine (i.e. as the number of cells becomes large), C_1 approaches the value of the correlation coefficient r, if the population is normal. Unfortunately, like the correlation ratio, C_1 depends on the form and fineness of grouping, but it gives an appreciation of the strength after the reality of an association has been established by the χ^2 method of Chapter IV, and does not depend on the order in which the groups are arranged. The contingency coefficient is not reliable when there are fewer than about sixteen cells, and, as in calculating χ^2, cells with very few entries should be avoided as far as possible; the constant is thus only suitable for large samples. Subject to this precaution, the finer the grouping the better, except that the computation becomes very laborious with a very large number of cells. For the smallpox data of Table 4.5,

$$\chi^2 = 196{\cdot}33, \quad N = 1\,689,$$

so

$$\phi^2 = 0{\cdot}116\,2 \quad \text{and} \quad C_1 = \sqrt{\frac{0{\cdot}116\,2}{1{\cdot}116\,2}} = 0{\cdot}32.$$

It would be instructive to combine groups of Tables 7.1–7.5 in various ways so as to form contingency tables with different numbers of cells and thus to see how the contingency coefficients approach the correlation coefficients.

The equivalence of the constants of this section to the correlation coefficient depends on the assumption that the variates are distributed normally, but that is no real limitation, for when they can only be expressed qualitatively we may reasonably imagine a quantitative description on such a scale that its distribution is normal.

CHAPTER X

THE FURTHER THEORY OF ERRORS AND PRINCIPLES OF EXPERIMENTAL ARRANGEMENT

Theory of Errors

10.1. The sampling distributions used in the applications of the theory of errors in Chapters II to V have all been based on the assumption that the individuals in a sample are independent. When, however, the variation is heterogeneous and the extent to which the sources of variation are represented is not left to chance, the individuals are not independent and the simple theory does not apply. This condition arises in Table 6.1 (p. 127) where each shrub is represented equally and the ovaries from one shrub are related. The same general theory of statistical inference applies, since it does not depend on any assumptions of independence, but the sampling distributions and standard errors have to be modified; we shall discuss these modifications and some of their consequences in the next three sub-sections.

Random Sampling from Unlimited Field

10.11. Considering the sample of ovaries of Table 6.1 (p. 127) as representing an infinite population of shrubs, let us estimate the standard error of the mean. Using the simple theory, we should have estimated the standard deviation from the distribution in the "totals" column, and from Table 6.2 we see that this gives as an estimate of standard error the value

$$\pm\sqrt{\frac{5{\cdot}385}{1\,000}} = \pm 0{\cdot}073\,4.$$

We know this to be incorrect, since the ovaries are not independent. However, the 10 shrub-means are independent and may be regarded as a small random sample from the infinite population of shrub-means. From Table 6.2, the variance between shrub-means after allowing for weighting is 2·614 92 and the standard error based on 9 degrees of freedom is

$$\pm\sqrt{\frac{2{\cdot}614\,92}{10}} = \pm 0{\cdot}511\,4.$$

Subject to the limitations of a small sample, this estimate is correct, and is about seven times the incorrect estimate.

It may be seen how this difference arises when allowance is made for the fact that the standard error of the mean is made up of two parts, one due to the true variance between shrubs, σ_s^2, and the other due to the variance within a shrub, σ_r^2. If there are m shrubs and n ovaries per shrub, the first source of variation is only represented by m individuals and the second by mn, and therefore we have

$$\textit{Standard Error of Mean} = \sqrt{\frac{\sigma_s^2}{m} + \frac{\sigma_r^2}{mn}}, \quad . \quad . \quad (10.1)$$

and from equation (6.3), p. 129, we see that the best estimate of this is $\sqrt{v_s/nm}$. The simple theory assumes that both sources of variation are sampled independently mn times. If we took 1 000 ovaries one at a time from the population of shrubs, leaving it to chance from which shrubs they came, this condition would be satisfied and we should have

$$\textit{Standard Error of Mean} = \sqrt{\frac{\sigma_s^2 + \sigma_r^2}{mn}} \quad . \quad . \quad (10.2)$$

If the estimates of σ_s^2 and σ_r^2 given in section 6.2 are substituted in (10.2) the result differs only slightly from the first estimate of the standard error given in this chapter.

Complex samples of the kind exemplified by the 1 000 ovaries, with 100 from each shrub, are not uncommon, and standard errors may easily be estimated if the sample can be divided into a number of equal, independent sub-samples or units like the shrubs. For example, the 500 pairs of brothers mentioned in section 3.1 are 500 independent units to which the sampling theory applies. To find the standard error of the mean character of the sample of cotton hairs collected in tufts as described in section 3.1, it would be necessary to measure the mean character separately for each of the reduced tufts, say 64 in number, and then calculate the standard error of the grand mean of these 64 tuft means. Sometimes, the sub-samples themselves are not independent, but have to be grouped into yet larger units, and there may be a whole hierarchy of units, only the largest of which are independent. Thus, eight tufts of 100 cotton hairs may be taken from each of eight bales to sample a

consignment of cotton, and since only the bales are independent, we must reduce our 6 400 hairs to a small sample of eight bale means.

The same considerations apply to means of observations in a time (or space) series, say to the mean annual rainfall obtained from the records for a number of years. The total variations are due to residual deviations and to a gradual trend, which may perhaps be expressible by a polynomial equation with a few constants; the trend corresponds to the shrub or tuft variations in the above examples, and is not sampled as many times as there are years. Indeed, it is difficult to say how many times it *is* sampled, and exactly what contribution it makes to the standard error of the mean, and we can only be sure that if the secular trend is at all important, the simple standard error given by $\sigma/\sqrt{N}$ is a serious underestimate of the true one.

The situation may be described in another way by stating that when the variation is heterogeneous, and the individuals are not independent, the variations between the individuals in a sample are not of the same kind as those between samples. Consequently, the simple sampling theory which estimates the variance between samples from that within a sample, cannot apply. This general situation may arise in several ways. For example, a laboratory that conducts routine determinations of the chemical or biological qualities of materials may attempt to estimate its errors by doing replicate tests. If the replicates are always done in parallel, going through the various preparations and treatments together and on the same day, they are not subject to the same errors as replicates done on different days and subjected to processes of determination that are independent from the beginning. Consequently, standard errors calculated from the parallel replicates are not valid estimates of the error in the comparisons of independent determinations made on different materials on different days. It is essential to ensure that the replicates on one material should be subject to the same variations as the determinations of the different materials to be compared. It may be suggested here that where much routine sampling or testing of a particular kind is done, replicate independent samples or tests of a number of populations or materials may provide a good combined estimate of the error of that kind of determination. For example, the replicates may be duplicate determinations, and the variance may be estimated from the equation for pairs in section 5.1. Before doing this, however, it is well to conduct an

investigation to see that the estimates of variance provided by the separate sets of replicates do not differ among themselves by an amount greater (or less) than may be attributed to random errors. This may be done by Neyman and Pearson's test, section 5.5, if there are only a few sets of replicates; if there are many sets, an approximate test is provided by seeing that the variance of the separate estimates of variance is not significantly greater than $2v^2/N$ (section 3.55), where v is the mean variance for all sets and N is the number of replicates per set. Another line of investigation is to test for the existence of a correlation between the variance and mean quality of each set.

Economy in Sampling

10.12. When the variation in a population is heterogeneous, the arrangement may have an influence on the precision attainable with a given size of sample. We shall discuss here the best way of distributing observations between and within groups when there are those two simple sources of variation.

If we use the notation of the previous section and let the total number of observations, mn, be N, the standard error of the mean given by equation 10.1 reduces to

$$\sqrt{\frac{1}{N}(n\sigma_s{}^2 + \sigma_r{}^2)}.$$

For a given number of individuals, N, this is least when n is smallest, i.e. when $n = 1$. That is to say, unless there are technical objections, for a constant number of observations the maximum accuracy in the mean is obtained when each one is selected separately and individually from a separate group; and in that case the standard error found in the ordinary way from the total variance is applicable. The reader will agree that this conclusion is merely common sense.

However, technical difficulties often intervene; for instance, we have mentioned in section 4.1 that in selecting cotton hairs individually there is a tendency to prefer the long ones, and to overcome this, hairs have to be taken in tufts.

Another modification is necessary when it takes more time to increase the number of groups than to increase the number of observations within a group. Smith and Prentice (1929), in obtaining soil samples for counts of cysts, took a number of "borings" of soil,

and then made several counts on each boring; if there were no other considerations, it would be best to take many borings and to make one count on each, but we must take account of the fact that it may take more time to obtain a boring than to make a count on one. If there are n counts on each of m borings, σ_s^2 is the variance between borings in the infinite population, σ_r^2 that of counts within a boring, t the time to make a count and kt the time to make a boring, we have for the total time to obtain the $N = mn$ counts,

$$T = mnt + mkt; \quad . \quad . \quad . \quad . \quad . \quad (10.3)$$

and the square of the standard error of the grand mean,

$$\sigma_M^2 = \frac{\sigma_s}{m} + \frac{\sigma_r^2}{mn} \quad . \quad . \quad . \quad . \quad . \quad (10.31)$$

Now from (10.3),

$$mn = \frac{T}{t} - mk,$$

and substituting this in (10.31) we obtain

$$\sigma_M{}^2 = \frac{\sigma_s{}^2}{m} + \frac{\sigma_r{}^2}{\left(\frac{T}{t} - mk\right)}.$$

In order to find the best way of distributing a constant time T between taking borings and making counts, we must find the value of m which makes σ_M^2 a minimum. Performing this operation in the usual way by making

$$\frac{\partial \sigma_M{}^2}{\partial m} = 0,$$

and substituting, we obtain

$$n^2 = k\frac{\sigma_r}{\sigma_s}{}^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (10.32)$$

Smith and Prentice found estimates for σ_s and σ_r to be 40·5 and 23·1 per cent. of the mean, and if we assume it takes five times as long to take a boring as to make a count, we find from (10.32) that the time is best employed when $n = 1{\cdot}3$; i.e. under such circumstances it would be better to increase the number of borings than to take more than two counts on each one. Examples of this sort arise

in other connections, and equation (10.32), which is only intended to give a rough guide, may be found useful. Similar considerations also arise in estimating other constants.

Sampling from Limited Field

10.13. In considering the sampling theory connected with the data of Table 6.1, we have regarded the ten shrubs as a random sample from an infinite population of shrubs. Sometimes, however, there may be only ten shrubs in the population under investigation, i.e. the field of sampling may be limited, although an infinite population of ovaries may be postulated. Then there are two possible methods of sampling.

In the first, the method of unrestricted random sampling, ovaries are taken entirely at random without any regard being paid to the shrub from which they come. Then, in a finite sample, the numbers of ovaries from the several shrubs are not necessarily the same and in general shrub variations contribute something to the sampling errors. The "totals" column of Table 6.1 represents this population of indiscriminately mixed ovaries from the ten shrubs, and the variance estimated from this provides the standard deviation from which the standard error of the mean has already been estimated as $\pm$ 0·073 4. The equation for this standard error is equation (10.2), except that when σ_s^2 is obtained from the sample as in section 6.2 it should be multiplied by $(m - 1)/m$; if the number of shrubs is limited, σ_s^2 is not *estimated*, it is determined. In the subsequent discussion we shall neglect this correction to equation (10.2).

The second method is termed by Bowley (1926) *selection in strata* and it consists in dividing the population into a number of parts or strata and in taking a sample of individuals at random except that the number taken from each stratum is proportional to the number present. In our example the ten shrubs are the strata and they may be presumed to be approximately equal in size. Then since each shrub is equally represented in the sample, as far as shrub variations are concerned the sample is an exact representation and sampling errors are due entirely to within-shrub variations. The within-shrub variance is 3·057, so the standard error of the mean of N ovaries sampled in strata is

$$\sqrt{\frac{3\cdot057}{N}}, \quad \text{or generally,} \quad \sqrt{\frac{\sigma_r^2}{N}}$$

In the language of section 3.7, the method of selection by strata is

$$100\,\frac{\sigma_s^2}{\sigma_r^2}\text{ per cent.}$$

more efficient than purely random selection; for the data of Table 6.1 this difference in efficiency is 76 per cent.

The method of sampling by strata has a wide application since most populations are limited in extent, even though they may have so many individuals that the conception of an infinite population is reasonable. For example, an agricultural plot can be divided into areas for sampling ears of corn, say; a town can be sampled in wards and streets in a sociological survey as suggested in section 3.1; or the products of a factory can be divided into strata representing the output of different machines, operatives or periods of time. The degree of superiority of sampling by strata to simple random sampling depends on the variance between—compared with that within—the strata, and there is an art in choosing the lines of division so as to make the ratio as large as possible. A knowledge of the sources of variation and of their relative importance such as that given by an investigation using the analysis of variance and by *a priori* technical knowledge is of great assistance. When there are no real variations between the strata, i.e. when $\sigma_s^2 = 0$, sampling by strata is at its worst, and even then it is as good as random sampling. It should be noted that unless there are at least two individuals from each stratum, σ_r^2, and hence the precision of the sample, cannot be estimated.

The Principles of Experimental Arrangement

10.2. All the foregoing considerations apply to the determination of a mean; but quite often we are interested, not in the absolute value of the mean, but in the extent to which the mean changes with some change in conditions (varied, perhaps, experimentally). From the point of view of such experiments, the variations are disturbing factors which the observer has to average out by taking large numbers of observations. If the variability is heterogeneous, however, it is often possible so to arrange the experiment that the group differences affect all the conditions or treatments equally, so that only the smaller residual variations contribute to the errors in the comparisons, giving an increased accuracy for the same number of observations. This is the principle on which experiments should be designed.

Harris, in the paper from which the data for Table 6.1 were taken, says that there were three series of ovaries:

A, ovaries from opened flowers which were shaken from the shrub;
B, ovaries from opened flowers which were apparently continuing their development at the time of sampling; and
C, ovaries which had completed their development to mature fruit.

He wished to discover if series C was different from series B and (more particularly) from A, and compared among other things the mean numbers of ovules per ovary.* Now if each series had been taken at random from different lots of shrubs, the standard error of each mean would have been calculated from the variance between trees, and would have been $\pm 0{\cdot}511\,4$, as calculated in section 10.11. Actually, however, each series was taken from the same shrubs, and by comparing corresponding means, the shrub variation was eliminated. For the purpose of such a comparison, the standard error of the grand mean is obtained from the smaller intra-shrub variance, and is that given in section 10.13 for selection by strata, viz.

$$\pm\sqrt{\frac{3{\cdot}057}{1\,000}} = \pm 0{\cdot}055\,29,$$

or about one-ninth of the standard error for a random collection of ovaries from two series of independent trees. This increase of precision may be expressed by the fact that to obtain the same accuracy, the number of ovaries from two independent series of shrubs would have to bear to the number for the restricted arrangement the ratio

$$\frac{(0{\cdot}511\,4)^2}{(0{\cdot}055\,29)^2}$$

or about 86: 1 (assuming 100 ovaries to be taken from each shrub).

A simple instance of the advantage of arranging the experiment or collection of data so that large variations are eliminated is that of comparing the means of two parallel series of observations, and as an example we will compare the mean head breadths of termites from

* Actually, Harris compared the ovules per loculus, and we have taken the ovary as the unit because its distribution showed the heterogeneity of the variance better. However, Harris only dealt with ovaries with three loculi, so that it does not affect our argument.

nests 668 and 670 (Table 6.8, p. 137). First, ignoring the fact that corresponding pairs of breadths are from the same month by treating the 10 observations as though they were independent, we find according to section 5.3,

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{0{\cdot}191\,6}{0{\cdot}075\,9 \times 0{\cdot}632} = 4{\cdot}00.$$

Now, taking the five differences between corresponding pairs from the same month, we use the method of section 5.2 and find

$$t = \frac{d\sqrt{N}}{s} = \frac{0{\cdot}191\,6 \times 2{\cdot}236}{0{\cdot}081\,09} = 5{\cdot}28.$$

This increased value of t means that by comparing samples taken in parallel months, the significance has been increased.

This fact may be expressed algebraically for the simple case; if there are two series of quantities, x_1' and x_2', both measured directly as deviations from their grand means, the sum of the squares of their differences is

$$S(x_1' - x_2')^2 = Sx_1'^2 + Sx_2'^2 - 2Sx_1'x_2'.$$

When we wrote a similar equation in section 6.1, we assumed the variates to be independent and the product term to be zero, but if the variates are not independent, we see from equation (7.4), p. 165, that

$$Sx_1'x_2' = r\sqrt{Sx_1'^2 Sx_2'^2}.$$

Substituting in the above equation, dividing by $(N - 1)$ and writing σ with an appropriate suffix for the standard deviation, we find

$$\sigma_{x_1 - x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2r\sigma_{x_1}\sigma_{x_2}.$$

From this, putting $r = 0$ we obtain the formula for the standard error of a difference between two independent means; but in pairs of parallel series like the head breadths of termites, r is positive and real and the standard error of the difference is thus reduced. If on the other hand r were negative, the standard error of the difference would be increased.

The Further Analysis of Variance

10.3. The whole question of the arrangement and interpretation of experiments is closely bound up with the analysis of variance, as we have shown in section 6.6 Usually there are more than two superimposed factors and the analysis may be complex. For example, when considering the variations in head breadths of termites of Table 6.8, we did two simple analyses into variations between and within nests, and between and within months. We see, however, that the nest and month variations are superimposed, and one analysis should separate the variations due to nests, months and residual random causes. Such an analysis will be developed in this section.

Since there is a real variation between nests, and since each month is represented equally in each nest, we may eliminate the nest variations by expressing the breadths as deviations from their nest means as in Table 10.1. We may now analyse the variance of these deviations into two parts, between and within months; this is done in Table 10.2. The "total" sum of squares is, of course, the "within a nest" sum of Table 6.9, and arises from 20 degrees of freedom. The "between months" sum is the same as before (the elimination of the nest means has made no difference to the differences between the months), and the degrees of freedom are still 4; the final residual, with 16 degrees of freedom, is thus very small. We now test for association with the months by finding the z for the variances; it is 1·16, and for $n_1 = 4$, $n_2 = 16$ lies well beyond the 0·01 point. Thus, by taking account of and eliminating the nest variations, we have established the month of year effect which previously appeared to be of doubtful significance.

There is no reason why we should not start with the deviations from the monthly means, and analyse their variance into two parts, between and within nests. Such a process would give the same between-nest variance as Table 6.9, and the same residual within a nest as is given in Table 10.2 for within a month. This residual is common, and represents random deviations that cannot be eliminated. The significance of the nest variance then becomes very much greater even than before, for

$$z = \tfrac{1}{2}[\log_e 0{\cdot}034\,36 - \log_e 0{\cdot}002\,31] = 1{\cdot}35,$$

and $n_1 = 4$ and $n_2 = 16$. We have thus analysed the total variations into three parts, between nests, between months and a residual;

TABLE 10.1

Mean Head Breadths of Termites (Small Soldiers)

(Deviations in mm. from nest means)

Nest Number	668	670	672	674	675	Means
November ..	− 0·061 2	− 0·046 8	+ 0·006 4	+ 0·025 6	− 0·066 4	− 0·028 48
January ..	− 0·002 2	+ 0·077 2	+ 0·059 4	− 0·033 4	+ 0·103 6	+ 0·040 92
March	+ 0·040 8	+ 0·087 2	+ 0·054 4	+ 0·093 6	+ 0·110 6	+ 0·077 32
May	+ 0·038 8	+ 0·031 2	− 0·001 6	+ 0·023 6	− 0·035 4	+ 0·011 32
August	− 0·016 2	− 0·148 8	− 0·118 6	− 0·109 4	− 0·112 4	− 0·101 08

TABLE 10.2

Analysis of Variance

(Deviations of Head Breadths of Termites)

Source of Variations	Sum of Squares	Degrees of Freedom	Variance
Between months	0·094 046	4	0·023 51
Within a month	0·036 921	16	0·002 31
Total (within a nest) ..	0·130 967	20	—

each of these can be given a place in a larger table of analysis with four rows.

The algebraic relations may be expressed by the equation,

$$(x - \bar{x}) = (\bar{x}_s - \bar{x}) + (\bar{x}_t - \bar{x}) + d, \quad . \quad . \quad (10.4)$$

where x is an individual reading for the sth row and the tth column,*

$\bar{x}_s$ is the mean for the sth row,

$\bar{x}_t$ is the mean for the tth column,

$\bar{x}$ is the grand mean, and

$d = x - \bar{x}_s - \bar{x}_t + \bar{x}$ is the residual deviation.

Then squaring, and summing for all rows ($\underset{s}{S}$) and columns ($\underset{t}{S}$), we have

$$\underset{s}{S}\underset{t}{S}(x - \bar{x})^2 = \underset{s}{S}\underset{t}{S}(\bar{x}_s - \bar{x})^2 + \underset{s}{S}\underset{t}{S}(\bar{x}_t - \bar{x})^2 + \underset{s}{S}\underset{t}{S}d^2,$$

since the sums of the product terms come to zero. These are the sums of squares that go into the analysis table. If there are n_t rows and n_s columns (i.e. n_s observations in each row and n_t in each column, total $N = n_s n_t$), the above equation becomes

$$\underset{s}{S}\underset{t}{S}(x - \bar{x})^2 = n_s \underset{s}{S}(\bar{x}_s - \bar{x})^2 + n_t \underset{t}{S}(\bar{x}_t - \bar{x})^2 + \underset{s}{S}\underset{t}{S}d^2 \quad . \quad (10.5)$$

Expressed in words, this equation states that the sum of squares of deviations of individual observations from the grand mean equals the sum of squares of deviations of row means multiplied by the number of readings in each row, plus the sum of squares of deviations of column means multiplied by the number of readings in each column, plus the sum of squares of residual deviations.

These are entered in Table 10.3, and symbols are written for the variances. If ${\sigma_s}^2$, ${\sigma_t}^2$ and ${\sigma_r}^2$ are the squares of standard deviations of row, column and residual variations in the infinite population, we obtain as in equations (6.3)

$$\left.\begin{aligned} v_s &\longrightarrow n_s{\sigma_s}^2 + {\sigma_r}^2, \\ v_t &\longrightarrow n_t{\sigma_t}^2 + {\sigma_r}^2, \\ v_r &\longrightarrow {\sigma_r}^2. \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (10.6)$$

After the significances of the sources of variation have been established by the z test carried out on the values of v, it is advisable

* The conception is generalised by writing row for month and column for nest.

to calculate these estimates of σ^2, the "corrected variances" as we may call them, as measures of the relative importance of the different sources. The values of v depend on the number of observations per row or column, i.e. on arbitrary features of the sample and so are of no value in themselves for measuring the objective properties of the population. If a large number of observations were taken individually at random from a population comprised of many rows and columns, of which the few rows and columns treated in the analysis are a random sample, the variance between these individuals would tend to equal the value $\sigma_r^2 + \sigma_s^2 + \sigma_t^2$. If any factor were eliminated as a source of variation, this variance would be reduced

TABLE 10.3

ANALYSIS OF VARIANCE

Source of Variations	Sum of Squares	Degrees of Freedom	Variance
Rows	$n_s \underset{s}{S}(\bar{x}_s - \bar{x})^2$	$n_t - 1$	v_s
Columns	$n_t \underset{t}{S}(\bar{x}_t - \bar{x})^2$	$n_s - 1$	v_t
Residual	$\underset{s}{S}\underset{t}{S}d^2$	$N - n_s - n_t + 1$	v_r
Total	$\underset{s}{S}\underset{t}{S}(x - \bar{x})^2$	$N - 1$	—

by the amount of the corresponding variance. For the head breadth of termites,

$$\sigma_r^2 \to 0{\cdot}002\,31 \text{ (residual)},$$

$$\sigma_s^2 \to \frac{0{\cdot}023\,51 - 0{\cdot}002\,31}{5} = 0{\cdot}004\,24 \text{ (months)},$$

$$\sigma_t^2 \to \frac{0{\cdot}034\,36 - 0{\cdot}002\,31}{5} = 0{\cdot}006\,41 \text{ (nests)};$$

and the variations due to months and nests are seen to be comparatively important. Sometimes, however, when there are many rows or columns, an unimportant variation may be highly significant. To interpret these variances in terms of normal frequencies, reference may be made to section 2.52.

The estimates of σ_s^2 and σ_t^2 are subject to errors of random sampling, but since these corrected variances are differences between

two independent estimates of variance, their sampling distributions are composite. We shall make no attempt to determine them here, since all the tests of significance that are usually necessary can be performed on the values of v.

In computing the residual sum of squares for Table 10.3, it is troublesome to find the actual residual deviations, d, as we have done above, and provided the arithmetic is carefully checked, it is sufficient to find the sums of squares for the rows, columns and total, and to obtain the residual by subtraction. The procedure is facilitated by extending the methods of section 6.3, performing the following operations:

(1) Sum the squares of the individual observations,
(2) Sum the squares of the row totals and divide the sum by the number of individuals in each row,
(3) Sum the squares of the column totals and divide the sum by the number of individuals in each column,
(4) Square the grand total and divide by the grand total number of observations.

Then it may easily be shown that the sums of squares required for Table 10.3 are:

For rows	..	the result of operation (2) minus that of (4),
For columns	..	(3) — (4),
For the total	..	(1) — (4).

The observations may all be measured from some arbitrary origin, e.g. 2·200 mm. for the head breadths of termites, so that the quantities to be squared have only three or four significant figures, and the computations may all be performed easily with the aid only of a table of squares.

10.31. A statistical study of the variations in a population is useful for establishing a sampling scheme, as suggested in section 10.13, and in industrial practice for suggesting in what directions it would be profitable to attempt to control quality. It is usually possible to collect a sample in which the sources of variation are represented in some balanced way as in Table 6.8 and to carry out an analysis of variance. This represents the type of control by selection and arrangement referred to in the Introduction to this book. Analyses

carried out in this connection may be very complicated, and it will be instructive to deal with the following example in detail.

The data in Table 10.4 are from a paper by Gould and Hampton (1936).* From a single "pot," about eighteen cylinders of spectacle glass are made on any one day or "journey." Two pots are heated together in the same furnace and glass may be made on several consecutive days from the same pair of pots, forming a "run." The quality measured, or variate, is the mean number of seed per unit area of glass, and this is given in Table 10.4 for three cylinders from each pot—the third, tenth and sixteenth in order of manufacture—and for the first five journeys in each of four runs. There is no significance in the numbering of the pots, and the runs are independent, being separated by several weeks and sometimes referring to different furnaces. The separate figures are given in the table, and various totals† have also been computed.

Now each kind of manufacturing unit—cylinder, pot, journey or run—is a potential source of variation, since the causes that control quality may be consistently different for each unit. Further, since the journeys are consecutive days and the cylinders are in order of manufacture, there may be trends in quality. Clearly, the variations in density of seed shown in Table 10.4 are the result of the superimposition of variations from many possible sources. Precisely what are these sources? Which have statistically significant effects? What is their relative importance? To answer these questions we shall analyse the variations in Table 10.4 completely.

It would be possible to do this analysis in one step writing down equations analogous to equations (10.4) and (10.5), but the notation would be complicated and the process would be very difficult to follow. We shall break Table 10.4 up into a number of simpler tables of kinds that have already been dealt with, and analyse these step by step. Readers are warned that they will find this analysis very difficult to follow unless they have "at their finger-tips" the simpler analyses dealt with so far.

First let us consider the fifteen readings in one pot-run, say the first pot and first run. This is statistically similar to Table 6.8

* From Table IV of their paper. We give here only four of the five runs given there, so as to make the analysis easier to follow. It might lead to some ambiguity in the argument if five runs and five journeys were retained.

† These totals are not total frequencies as given, say, in Table 6.1. They are the sums of the qualities of the individuals included in the totals.

TABLE 10.4

MEAN NUMBER OF SEED PER UNIT AREA

Run	Journey	Pot 1			Pot 2			Totals		
		Cylinder			Cylinder			Pot 1	Pot 2	Both Pots
		3	10	16	3	10	16			
1	1	47	56	100	52	61	88	203	201	404
	2	55	89	93	49	62	97	237	208	445
	3	35	57	56	34	60	72	148	166	314
	4	78	67	113	47	93	118	258	258	516
	5	33	40	128	16	29	130	201	175	376
	Totals	248	309	490	198	305	505	1 047	1 008	2 055
2	1	52	66	36	65	80	40	154	185	339
	2	21	61	49	122	97	79	131	298	429
	3	31	39	25	45	54	72	95	171	266
	4	43	72	52	109	120	80	167	309	476
	5	37	51	67	67	85	63	155	215	370
	Totals	184	289	229	408	436	334	702	1 178	1 880
3	1	50	61	60	75	139	130	171	344	515
	2	33	27	49	46	58	63	109	167	276
	3	24	39	24	15	33	39	87	87	174
	4	18	18	43	22	16	19	79	57	136
	5	28	42	28	27	19	22	98	68	166
	Totals	153	187	204	185	265	273	544	723	1 267
4	1	24	34	43	46	66	24	101	136	237
	2	24	49	42	40	117	105	115	262	377
	3	21	21	51	30	28	34	93	92	185
	4	21	69	48	36	64	53	138	153	291
	5	76	48	42	39	60	78	166	177	343
	Totals	166	221	226	191	335	294	613	820	1 433

TABLE 10.5

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
(*a*) *Within Pots*			
(1) Between cylinders..	22 132·53	16	1 383·28**
(2) Between journeys ..	37 916·66	32	1 184·90**
(3) Residual within pots	18 398·14	64	287·47
(4) Total	78 447·33	112	—
(*b*) *Between Pots*			
(5) Between runs ..	13 679·90	3	4 559·96
(6) Residual between pots	10 099·57	4	2 524·89**
(7) Total	23 779·47	7	3 397·07**
(*c*) *Between Cylinders*			
(8) Common to all runs	9 132·88	2	4 566·44
(9) Common to both pots in run less (8)	11 532·72	6	1 922·12
	20 665·60	8	2 583·20**
(10) Specific to pot ..	1 466·93	8	183·37
(11) Total	22 132·53	16	—
(*d*) *Between Journeys*			
(12) Common to all runs	9 684·00	4	2 421·00
(13) Common to both pots in run, less (12)	18 650·07	12	1 554·17
	28 334·07	16	1 770·88**
(14) Specific to pot ..	9 582·59	16	598·91*
(15) Total	37 916·66	32	—

(p. 137), and the variation may be divided into three parts: between cylinders with two degrees of freedom, between journeys with four degrees, residual with eight degrees, and there are fourteen degrees for the total. This process may be carried out for the eight pot-runs, and the sums of squares and degrees of freedom added, giving the data in the upper part of Table 10.5.

Now consider the pot means, $\frac{1047}{15}$, $\frac{1008}{15}$, $\frac{702}{15}$, etc. There are eight readings giving seven degrees of freedom and the sums of squares may be divided into two parts: between runs with three degrees and a residual within runs with four degrees. To obtain sums of squares comparable with those within pots, we should

TABLE 10.41

	Cylinder		
	3	10	16
Pot 1	49·6	61·8	98·0
Pot 2	39·6	61·0	101·0

multiply the sums of squares obtained from the pot means by the number of readings in a pot, 15, just as in equation (10.5) we multiplied the squares of the row and column deviations by the number in a row, n_s, or column, n_t. This is equivalent to stating that the squared deviations should be summed over all the original individuals. When this process is carried out, the results in section (*b*) of Table 10.5 are obtained.

The cylinder variations given in row (1) may be analysed further. The cylinder means for run 1 are in Table 10.41. This is like Table 6.8, and the variance may be analysed as in Table 10.3 into parts associated with pots (one degree), a cylinder variation common to both pots (two degrees) and a residual (two degrees) which represents an extra cylinder effect that is specific to each pot and is due to the deviations of the separate cylinder readings for each pot from the cylinder readings common to both pots. Similar analyses may be carried out for the four runs, and the sums of squares and degrees of freedom added. When this is done and the resulting sums of squares are multiplied by five, to make them

comparable with section (*a*) of Table 10.5, the results are as shown in rows (6), (8) and (9) together, and (10) of that table. It would be just as reasonable, statistically, to describe the residual in row (10) as an extra pot effect that is specific to each cylinder, and to show impartiality, we may call the source of variation an interaction between cylinders and pots. From a practical point of view, however, it seems more appropriate in this instance to concentrate attention on the cylinder variations.

Yet another subdivision of the cylinder effect common to both pots in a run (rows 8 and 9) is possible. The cylinder means common to both pots in each of the four runs are in Table 10.42. The total sum of squares between these means has 11 degrees of freedom

TABLE 10.42

	Cylinder		
	3	10	16
Run 1	44·6	61·4	99·5
Run 2	59·2	72·5	56·3
Run 3	33·8	45·2	47·7
Run 4	35·7	55·6	52·0

and may be divided into parts associated with runs (3 degrees), cylinders common to all runs (2 degrees) and a residual (6 degrees) representing a differential cylinder trend for each run but common to the two pots in the run, i.e. a cylinder-run interaction. The sums of squares of these items multiplied by 10 and the degrees of freedom are entered in rows (5), (8) and (9) of Table 10.5.

In a precisely similar way the journey-variations of row (2), Table 10.5, may be further analysed into the parts shown in section (*d*) of that table.

Before testing the variances in Table 10.5 for significance, it will be well to determine of what corrected variances they are estimates. Let the variances entered in the table be denoted by v with a subscript corresponding to the row of the table, i.e. v_1, v_2, etc., and let the corresponding corrected variances be denoted by σ^2 with a corresponding subscript. Thus, σ_1^2 is the variance that would be found between the cylinder means within each pot if an infinite

number of observations could be made on each cylinder and there were an infinite number of pots. Further, let the number of the original individuals contributing to the mean of each individual factor be n with an appropriate subscript. For example, n_1 is the number of readings per cylinder within each pot, $= 5$, n_9 is the number of readings per cylinder mean for each run, $= 10$, and so on. Then the values of the n's and the relations between the observed and corrected variances, expressed in the manner of equations (10.6), are as follows:

$$\begin{aligned} n_1 &= 5 & v_1 &\to n_1\sigma_1^2 + \sigma_3^2 \\ n_2 &= 3 & v_2 &\to n_2\sigma_2^2 + \sigma_3^2 \\ n_3 &= 1 & v_3 &\to \sigma_3^2 \end{aligned}$$

$$\begin{aligned} n_5 &= 30 & v_5 &\to n_5\sigma_5^2 + v_6 \\ n_6 &= 15 & v_6 &\to n_6\sigma_6^2 + \sigma_3^2 \end{aligned}$$

$$\begin{aligned} n_8 &= 40 & v_8 &\to n_8\sigma_8^2 + v_9 \\ n_9 &= 10 & v_9 &\to n_9\sigma_9^2 + v_{10} \\ n_{10} &= n_1 = 5 & v_{10} &\to n_{10}\sigma_{10}^2 + \sigma_3^2 \end{aligned}$$

$$\begin{aligned} n_{12} &= 24 & v_{12} &\to n_{12}\sigma_{12}^2 + v_{13} \\ n_{13} &= 6 & v_{13} &\to n_{13}\sigma_{13}^2 + v_{14} \\ n_{14} &= n_2 = 3 & v_{14} &\to n_{14}\sigma_{14}^2 + \sigma_3^2 \end{aligned}$$

We shall now proceed to discuss the derivation of these expressions.

The relations between v_1, v_2 and v_3 are the same as those between the variances of Table 10.3, except that we have combined the estimates of several tables. It will be noted that the cylinder and journey effects may vary from one pot or run to another and σ_1^2 and σ_2^2 may themselves be complex. The residual variance σ_3^2 is due partly to the fact that only a limited number of sections on each cylinder were examined for seed and partly to a real variation that could not be associated with any factor.

The relation between v_5 and v_6 follows from equation (10.6), and from the fact that after performing the analysis on the pot means, we multiplied the sums of squares by the number of individuals per pot. Thus, when the analysis is performed on the pot means, the n_s of equation (10.6) is the number of pots per run $= 2$; but we multiplied the sums of squares by the number of readings per pot, $= 15$, so in the above equation, σ_5^2 is multiplied by the product

of these which is n_5. The residual variance measured between the pot means is the sum of two parts; it is the corrected variance between pots, σ_6^2, plus the square of the standard error or variance due to the fact that the residual variations within a pot are only sampled by $n_6 = 15$ individuals; this added variance is $\sigma_3^2/15$. The variance v_6 is n_6 times this measured variance between pot means. The cylinder and journey effects do not contribute to the error in a pot mean, since they are expressed as deviations from the pot mean.

The relations between v_8 and v_9, between v_9 and v_{10}, between v_{12} and v_{13} and between v_{13} and v_{14} may easily be deduced in a similar manner; these variances have all been found from analyses like that in Table 10.3. The apparent variance v_{10} is contributed to only by the part of the cylinder trend that is specific to the pot, σ_{10}^2, and by the first residual, σ_3^2, and the relation given above is easily deduced. Similar remarks apply to the journey effect.

We are now in a position to test the variances given in Table 10.5 for statistical significance, using Fisher's z-test. If there is no cylinder effect, $\sigma_1^2 = 0$ and $v_1 = v_3$; the test for this effect consists in establishing the statistical significance of the difference between v_1 and v_3. Similarly, v_2 should be compared with v_3, v_5 with v_6, v_6 with v_3, v_8 with v_9 and so on. When these tests are carried out, the values of z corresponding to the variances marked with two asterisks are above the 1 per cent. point and that with one asterisk is between the 5 and 1 per cent. points; we shall presume that these show the existence of significant effects;* all the other values of z are below the 5 per cent. point.

The variances v_1 and v_2 are both significantly greater than v_3; if this had not been so, we should still have been justified in performing the further analyses of sections (*c*) and (*d*) of Table 10.5.

The variance v_5 is not significantly greater than v_6, so it is reasonable to combine these to give an improved estimate of the residual variance (*b*) between pots, based on 7 degrees of freedom; this is in row (7) of Table 10.5. Similarly, v_8 is not significantly greater than v_9 and it is reasonable to work on the conclusion that the cylinder effect varies from run to run and is measured by the combined variance based on 8 degrees. This combined estimate of v_9 is significantly greater than v_{10}. We see, however, that v_{10} is

* The one variance with one asterisk has a value of z very little below the 1 per cent. point.

not greater than v_3, and conclude that the cylinder effect does not vary from pot to pot within a run. Similar conclusions may be reached regarding the journey variances, except that v_{14} is greater than v_3. We are justified in combining rows (10) and (3) to obtain an improved estimate of v_3 based on 72 degrees of freedom; it is 275·90.

The significant effects and their corrected variances estimated from the apparent variances of Table 10.5 may be summarised as follows:

Differences between pots,

$$\sigma_6^2 \to \frac{2\,524{\cdot}89 - 275{\cdot}90}{15} = 149{\cdot}93$$

Cylinder effect varying from run to run,

$$\sigma_9^2 \to \frac{2\,583{\cdot}20 - 275{\cdot}90}{10} = 230{\cdot}73$$

Journey effect, part due to variation from run to run,

$$\sigma_{13}^2 \to \frac{1\,770{\cdot}88 - 598{\cdot}91}{6} = 195{\cdot}33$$

Journey trend, part due to variation from pot to pot within a run,

$$\sigma_{14}^2 \to \frac{598{\cdot}91 - 275{\cdot}90}{3} = 107{\cdot}67$$

Residual (unaccounted variation),

$$\sigma_3^2 \to \qquad\qquad 275{\cdot}90.$$

These estimates of corrected variance measure the relative importance of the several factors in the way described in section 10.3. The estimates are valid, however, only in so far as the third, tenth and sixteeenth cylinders are representative of the eighteen or so that may be made from a pot, and the first five journeys are representative of all the journeys possible in a run, in so far as there is any secular variation. For example, all the cylinders between the third and tenth could have fewer and those between the tenth and sixteenth could have more seed than the three measured, and this would be a source of variation entirely overlooked by our analysis, of which random errors take no account. We shall assume the representativeness of the cylinders and journeys.

All the above values of the true variances are estimates subject to random errors, since the cylinder and journey effects vary from run to run or from pot to pot, and the few that have provided the estimates are a random sample from an infinite population of trends obtained from an infinite population of runs. Had there been a significant cylinder effect common to all runs, it would not have been subject to random errors in quite the same way. However many runs there had been, there would have been but two degrees of freedom for v_8; the effect of a large number of pots is merely to increase n_8 in the equation

$$v_8 \to n_8\sigma_8^2 + v_9$$

and reduce the error v_9, thus reducing the effect of random errors in v_9 in estimating σ_8^2. Similar remarks apply to the journey effect.

We have developed the foregoing analysis by treating *means*—pot means, cylinder means, and so on. The computation is much easier, however, if corresponding *totals* are dealt with in the way outlined at the end of section 10.3. The general rule is first to square and add the various totals and divide by the number of individuals forming each total, producing a series of "adjusted sums." Then the sum of squares entered in Table 10.5 for any factor measured as a series of deviations from the means corresponding to a second factor is the "adjusted sum" for the first factor minus that for the second. As an example, let us find the sum of squares entered in Table 10.5, below rows (8) and (9). The first factor is represented by the run-cylinder totals and the second by the run totals (i.e. in the analysis, run-cylinder means were measured as deviations from run means for all cylinders). There are 10 observations per run-cylinder total and 30 per run total; the adjusted sums are:

$$(446^2 + 614^2 + 995^2 + 592^2 + \ldots) \div 10 = 401\,205{\cdot}70$$

and $$(2\,055^2 + 1\,880^2 + \ldots\qquad\qquad) \div 30 = 380\,540{\cdot}10,$$

and the sum of squares is the difference between these, viz.: 20 665·60.

Arrangements for Agricultural and Analogous Experiments

10.4. In no subject has the application of the principles we have just introduced for reducing the errors of comparisons of treatments or conditions been so complete as in that of agricultural experiment.

The actual statistical technique is capable of application to other subjects, and it illustrates well the analysis of complex variations; for these reasons as well as because of the practical importance of the question itself, we shall deal explicitly with the arrangement of plots for field trials. There is an example of an application to another subject in section 10.5.

It is well known that there are regional changes in natural fertility over areas of land, and that if varieties of corn (for example) are compared by sowing one large plot with each and measuring the yields, there is a possibility that any measured differences may be due to variations in soil fertility rather than to varieties of corn.

TABLE 10.6

YIELDS OF GRAIN IN GRAMMES

(Plots $7\frac{1}{2} \times 2\frac{2}{3}$ ft.)

	1	2	3	4	5
1	360	324	306	274	277
2	287	324	282	252	277
3	274	302	247	234	256
4	316	304	232	261	271
5	354	364	299	302	301
6	295	294	298	313	268
7	197	201	236	209	217
8	335	320	259	205	207
9	267	262	279	332	268
10	291	300	350	383	264

For that reason the necessity of repeating the comparisons over several plots so that the effect of soil variations can be assessed is well recognised. It is true, also, that these variations often exhibit trends so that if small plots are taken the average fertility of neighbouring ones tends to be alike; the art of planning an experiment lies in arranging that the varieties or conditions under investigation (we shall call them treatments) occur in neighbouring plots, so that the larger regional changes in fertility are eliminated from the comparison. It is desirable also that the standard error of the differences due to the uneliminated variations in fertility may be estimated so that the significance of any difference can be tested.

For the purposes of studying soil variations, many uniformity

trials have been conducted, and the data of Table 10.6, taken from a paper by Christidis (1931), are the results of one conducted at Cambridge. They are the yields of corn from fifty adjacent plots of $7\frac{1}{2} \times 2\frac{2}{3}$ feet treated uniformly, and harvested separately. If we wished to compare five treatments (say), we could distribute them at random over the fifty plots. In such case the standard error of the mean for any one treatment would be $\sqrt{v_r/10}$, where v_r is the residual variance between plots after treatments have been eliminated, and that of the difference between two means would be $\sqrt{2}$ times this; here the treatments are the same, and the residual variance for such an arrangement is the mean variance for the fifty yields and equals 95 940/49 = 1 958.

10.41. We might expect to eliminate the variations between rows, however, by ensuring that every treatment occurs once in each row, so that only the residual *within-row* variance contributes to the errors of the differences. Labelling the treatments A, B, C, D and E, such an arrangement as the following could be tried,

A B C D E
A B C D E
— — — — —

We have analysed the variance of the yields of Table 10.6 into three portions as in Table 10.3—variations between rows, between treatments arranged as shown and a residual within rows; the results are in Table 10.7. There are ten rows giving nine degrees of freedom, five treatments giving four degrees, fifty plots giving forty-nine degrees for the total and leaving thirty-six for the residual. We know, however, that the five treatments are all the same, and so may combine the treatment and residual sums of squares to give a *within-row* variance; we see that by eliminating the row variations, the variance is reduced from 1 958 to 1 218; i.e. a comparison based on about 12 replicates properly arranged with one of each treatment per row is as good as one based on about 20 perfectly random replicates. In practice, if the treatments differ, their significance is tested on the basis of the residual variance of 1 063; in this example the treatments variance (2 616) paradoxically seems greater than the residual. For the difference $z = 0{\cdot}45$, and being only just below the 5 per cent. point, is suggestive (although not strongly) of a real effect. This is because we have neglected a fundamental principle

—the plots are not randomly distributed within the restrictions imposed, and the treatment differences coincide with *column* differences. The arrangement would have been sound if we had settled the order of the plots in each row separately by chance (say by drawing lettered tickets from a hat), and the residual variance (1 063 in Table 10.7) would not have been very different from the variance

TABLE 10.7

ANALYSIS OF VARIANCE OF CROP YIELDS

Source of Variations	Sums of Squares		Degrees of Freedom		Variance	
Between rows ..	47 201·6		9		5 245	
Between treatments ..	10 462·6	48 738·4	4	40	2 616	1 218
Residual	38 275·8		36		1 063	
Total	95 940·0		49		1 958	

within rows (1 218); the reader is recommended to try it out on the data.

Other systematic arrangements have been devised to overcome such obvious defects as shown by the one just condemned, and of them the "chessboard" had been very popular. This places every treatment as close to every other as possible, and an arrangement for seven treatments in rows of five is here shown.

A B C D E
F G A B C
D E F G A
B C D E F
— — — — — .

It will be seen that the treatments may be assigned in the same order, row after row, and yet they are all represented in every column. If there are as many treatments as plots per row, each one may be two places to the right of its position in the previous row, viz.

A B C D E
D E A B C
B C D E A
E A B C D
C D E A B .

The usual method of treating the results is to divide them into groups containing all the treatments in the order in which they occur; in the seven treatments example above, the first group would contain the first row and the first two plots of the second, the second would contain the last three plots in the second row and the first four in the third, and so on; for five treatments, the groups would be rows. Then the total variance is analysed into three parts—between groups, between treatments and a residual from which the standard error of a difference may be calculated. The danger of column variations giving spurious effects is eliminated, but such arrangements do produce regular, if more complex patterns, and there is a remote chance of the soil varying periodically and more or less in step with the pattern. Further, the arrangement does not make the best possible comparison, for, taking the set of five treatments, the groups are rows, and the residual variance (which for Table 10.7 is 1 218) includes column differences, although they do not enter into the comparisons. The estimates of errors in such systematic arrangements are thus not valid.

Restricted Random Arrangements of Plots

Random Groups

10.42. For large numbers of treatments it is better to have groups of as many plots as there are treatments, arranged in clumps rather than in rows, and to assign the position of each treatment within the group by chance. The random arrangement of five treatments in the rows of Table 10.6 is a special case of this, and the analysis of the yields is the same as in Table 10.7.*

The Latin Square

The application of the Latin square arrangement to field experiments is due to Fisher (1936).† If there are n treatments to be compared, a block of plots having n rows and n columns is marked out, and the treatments are distributed at random, with the restriction that each one must occur once in each row and once in each column. The field of Table 10.6 has 5×10 plots, and so is suitable

* The results may be written in a table of rows and columns, each row referring to a group and each column to a treatment; then the terms of Table 10.3 apply directly.

† The first edition of this book appeared in 1925.

for the formation of two Latin squares—we give an example below. The variance of such an arrangement may be fairly simply analysed into five parts: (1) between 2 blocks (1 degree of freedom), (2) between 10 rows (the deviations are measured from 2 block means, giving 8 degrees of freedom), (3) between columns (we regard the 10 columns of the 2 blocks as contributing separately to the variance, giving 8 degrees of freedom), (4) between treatments (4 degrees of freedom) and (5) a residual with 28 degrees of freedom, giving a total of 49 degrees. Equation (10.4) may be extended to

	1	2	3	4	5
1	B	E	A	D	C
2	D	A	E	C	B
3	E	D	C	B	A
4	A	C	B	E	D
5	C	B	D	A	E
6	D	B	C	A	E
7	A	C	E	D	B
8	B	A	D	E	C
9	E	D	B	C	A
10	C	E	A	B	D

give equation (10.7), where the same notation is used, the suffixes r referring to blocks, s to rows, t to columns and u to treatments.

$$(x - \bar{x}) = (\bar{x}_r - \bar{x}) + (\bar{x}_s - \bar{x}_r) + (\bar{x}_t - \bar{x}_r) + (\bar{x}_u - \bar{x}) + d. \quad (10.7)$$

The rows and columns are measured as deviations from their block means, the blocks and treatments from the grand mean. When this is squared and summed, because of the restriction that each treatment is represented equally in each row and so on, the product terms become zero, and we have equation (10.8), which is parallel to (10.5) and may be entered in a table of analysis of variance.

$$S(x - \bar{x})^2 = n^2 \underset{r}{S}(\bar{x}_r - \bar{x})^2 + n \underset{s}{S}(\bar{x}_s - \bar{x}_r)^2 + n \underset{t}{S}(\bar{x}_t - \bar{x}_r)^2 + 2n \underset{u}{S}(\bar{x}_u - \bar{x})^2 + Sd^2, \quad . \quad (10.8)$$

where n is the number of rows per block.

These terms are the sums of squares of deviations of the various means from the block or grand mean, multiplied by the number of observations contributing to those various means. The actual arith-

metic is far less complicated than the algebra, and if the reader will work through an example, he will soon gain confidence.

When there is an awkward number of treatments (like 6), so that the means are not correct when calculated to one or two decimal places, the following arithmetical scheme, extended from that in section 10.3, is convenient:

(1) Square every plot yield and sum,
(2) Find the total yields for the blocks, square and sum them and divide by the number of plots per block,
(3) Find the total yields for the rows, square and sum them and divide by the number of plots per row,
(4) Repeat this for columns,
(5) Repeat this for treatments, and
(6) Square the grand total yield and divide by the total number of plots.

Then the terms in the following are equivalent to the corresponding ones in equation (10.8),

$$[(1) - (6)] = [(2) - (6)] + [(3) - (2)] + [(4) - (2)] + [(5) - (6)] + Sd^2.$$

The data of Table 10.6, with the above arrangement of imaginary treatments, have been reduced in this way, and the analysis is in Table 10.71. The treatment variance, although greater than the final residual, is not significantly so ($z = 0{\cdot}38$ while $z = 0{\cdot}50$ lies on the 5 per cent. point), and we will combine them and regard the mean variance 866 based on 32 degrees of freedom as the residual. The row and treatment variances are both greater than the residual (the z's lie above the 5 per cent. point), and the arrangement has reduced the residual variance to 866 (i.e. 9 replicates so arranged are as good as 20 at random).

The block variance may not be compared with the residual, because the rows and columns are not common to both blocks, and their variations contribute to the block differences.

For comparing treatment means, the standard error of each mean is $\sqrt{866/2n}$, where $2n$ is the number of plots per treatment. In making such comparisons of several means, due regard must be paid to the general considerations advanced in section 3.33.

It may be mentioned in passing that Latin square arrangements

are useful in many fields of experiment and have also been used for separating sources of variation not subject to experimental control, as was done in section 10.31. The "treatments," the rows and the columns may represent three superimposed factors, and there may be several squares, each with different individual representatives of the factors. Then the three corrected variances are of more interest than the means for the "treatments" and their standard errors.

Every separate experiment must be arranged in its own, independently and randomly chosen Latin square. To form a square,

TABLE 10.71

ANALYSIS OF VARIANCE OF CROP YIELDS

Source of Variation	Sums of Squares		Degrees of Freedom		Variance	
Blocks	3 698·0		1		3 698	
Rows	43 503·6		8		5 438	
Columns	21 042·4		8		2 630	
Treatments ..	6 452·6	27 696·0	4	32	1 613	866
Residual	21 243·4		28		759	
Total	95 940·0		49		1 958	

a number of cards may be lettered, one for each treatment, and shaken up in a hat. Then the letters may be assigned to the position in a row in the order in which the corresponding cards are drawn. The whole procedure may be repeated for the second and subsequent rows. Proceeding in this way, however, it will usually be found that after three or four rows, one or two letters appear more than once in a column. Whenever this happens, the row should be re-drawn until the restriction of one treatment per column is satisfied. In six-by-six or larger squares, a good deal of adjustment may be necessary and this may allow some scope for the unconscious introduction of a slight degree of system in the arrangement, which may tend to repeat itself if one experimenter has to make many squares. After a Latin square has been designed, however, the rows and columns may be rearranged in some random order fixed by drawing numbered

cards from a hat, and if this is done it is hardly likely that any serious lack of randomness will survive in the scheme. However, workers who have to form many Latin squares are advised to use such methods as described by Yates (1933*a*).

Use of Controls

10.43. One common method of comparing a number of experimental treatments when there are disturbing variations is to repeat one of the treatments, a "control," with each of the others, under conditions as similar as possible. Then, the differences between the treatment and control values are measured, and these differences are used to compare the treatments. In order that the precision of such an experiment may be measured, the treatment-control pairs must be replicated, and the relative positions of the treatment and control within each pair must be fixed independently and at random.

For example, let us suppose we wish to compare five treatments A B C D E on the plots referred to in Table 10.6, repeating A as a control with each of the others in a contiguous plot in the same row. Then we can only use four columns and the arrangement may be like the following:

Col. 1	*Col.* 2	*Col.* 3	*Col.* 4
B	A	A	C
A	D	E	A
A	B	C	A
D	A	E	A
etc.	etc.	etc.	etc.

The differences between the treatments and controls may be found; if the variance within the treatments of these differences is v and the number of replicates of each treatment is n, the standard error of the mean difference between any treatment and the control is $\sqrt{v/n}$, while that of the mean difference between any two of the treatments B to E is $\sqrt{2v/n}$.

Since the data of Table 10.6 are the result of a uniformity trial, v may be estimated as twice the variance within the pairs, and from the formula at the end of section 5.1 (p. 111), we find that $v = 737{\cdot}6$. The standard error of the difference between any two of the four treatment effects is therefore $\sqrt{1\,475{\cdot}2/n}$ and $8n$ plots have been used. The standard error of the difference between treatment A

and the other four is $\sqrt{737 \cdot 6/n}$. If the four treatments and control were distributed among $8n$ plots in the random arrangements, there would be $8n/5$ replicates and the standard error of the difference between any two treatments would be

$$\sqrt{\frac{2 \times 1958 \times 5}{8n}} = \sqrt{\frac{2447 \cdot 5}{n}}.$$

The method using a control is, in this instance, more efficient than the random method. On the other hand, similar calculations show the Latin square method to be more efficient than either except for comparing treatment A with B, C, D or E.

The use of a control may provide very accurate comparisons if it is technically convenient to bring each treatment and its control into very close proximity. This is achieved in Beavan's "Half-drill strip" method in which adjacent long narrow strips are assigned to the control and treatment in pairs. Often, the arrangement is such as to sandwich two strips of treatments between two of controls. This practice is to be deprecated, however, as it violates the principle of randomness; within each pair of strips the relative positions of the treatment and control should be decided by chance.

Multiple Factor Experiments

10.5. A multiple factor or *factorial* experiment is one in which several factors are varied together. For example, in an experiment arranged to investigate the effect of an ingredient of the size mixture used for protecting cotton warps in weaving, there were four treatments in which there were two forms of this ingredient, *A* and *B*, and each was used in the size in two concentrations: 1 and 2 units. The treatments were:

(i) 1 unit of ingredient *A*
(ii) 2 units of ingredient *A*
(iii) 1 unit of ingredient *B*
(iv) 2 units of ingredient *B*.

Further, the experiment was done twice on two separate qualities of yarn, *X* and *Y*. Thus there were three factors: form of ingredient, quantity of ingredient and quality of yarn. The quality measured on the warps was the rate at which the warp threads broke during weaving, the rate being expressed as the number per 10 000 picks.*

* One "pick" corresponds to the insertion of one weft thread.

Although the three factors were varied together in the same experiment, the arrangement is balanced so that both ingredients were present in both quantities and were used on both yarns. Consequently, the deviations of the two mean breakage rates for ingredients A and B from the grand mean are unaffected by anything but the forms of ingredient and errors. These mean deviations are said to measure the *main effect* of ingredients. Similarly, the main effects of the quantities of ingredient and of the yarn qualities may easily be measured, each being undisturbed by the other two factors.

TABLE 10.8

WARP BREAKS PER 10 000 PICKS AND TREATMENTS

Loom.. ..	7	8	9	10
Period		*Yarn X*		
1	1·4 (iv)	5·3 (i)	7·4 (ii)	3·2 (iii)
2	2·2 (iii)	4·1 (ii)	8·8 (i)	1·6 (iv)
3	2·1 (ii)	2·2 (iv)	4·8 (iii)	4·7 (i)
4	2·4 (i)	2·3 (iii)	2·6 (iv)	4·3 (ii)
Loom	15	16	17	18
Period		*Yarn Y*		
5	2·8 (iii)	1·5 (i)	1·9 (ii)	2·0 (iv)
6	1·5 (iv)	1·2 (ii)	2·0 (i)	1·8 (iii)
7	2·5 (ii)	1·4 (iv)	2·4 (iii)	2·0 (i)
8	5·8 (i)	1·9 (iii)	1·7 (iv)	3·2 (ii)

It is possible, however, that the effect of varying the quantity of ingredient may be different for form A than that for form B, or alternatively, the difference in response of warp breakage rate to A and B may depend on the quantity of each present. If this is so, there is said to be an *interaction* between form and quantity of ingredient. There may be interactions between any pair of the three factors. If such interactions exist, it may also be that the effect of increasing the quantity, say, from 1 to 2 units may be different according to whether the ingredient is A on yarn X, A on Y, B on X or B on Y. If such an effect exists, there is said to be a *second-order interaction* between the three factors.

The main effects and interactions may be measured and tested by an analysis of variance on the results, and with this in view we shall deal with results of the above experiments in full.

There were four warps* of each yarn, one treatment being used for each warp. The warps of yarn *X* were woven in four looms, and the total weaving time was divided into four periods. Between the periods, the warps were interchanged among the looms in such

TABLE 10.9

ANALYSIS OF VARIANCE OF WARP BREAKAGE RATES

Source of Variation	Sums of Squares	Degrees of Freedom	Variance
(1) Yarns	17·701	1	17·701
(2) Periods	10·348	6	1·725
(3) Looms	36·763	6	6·127
(4) Treatments	28·978	6	4·830
(5) Residual	8·539	12	0·712
(6) Total	102·329	31	—
Main Effects			
(7) Quantity	5·120	1	5·120
(8) Ingredient	17·111	1	17·111
Interactions			
(9) Quantity—Ingredient ..	0·046	1	0·046
(10) Quantity—Yarn	0·320	1	0·320
(11) Ingredient—Yarn	6·302	1	6·302
(12) Quantity—Ingredient—Yarn	0·079	1	0·079

a way as to complete a Latin square arrangement with periods as rows and looms as columns. The warps of yarn *Y* were woven in separate looms and periods from *X*, in another Latin square. The arrangement and warp breakage rates are given in Table 10.8.

First let us analyse these data as we did the crop yields in the double Latin square of section 10.42, except that since we are considering the possibility of yarn-treatment interactions we will keep separate the four treatments in each block, giving eight duplicated treatments with six degrees of freedom. The analysis is in rows (1) to (6) of Table 10.9. Rows (2) to (5) would be obtained if the two Latin squares were analysed separately and the results com-

* A warp is a unit quantity of yarn sized uniformly.

bined. The variance due to yarns in Table 10.9 may not be compared with the residual term, since loom and period variations contribute to the yarn differences. The variance due to treatments in row (4) is significantly greater than the residual, so we may analyse the treatment effects further.

The yarn effects may be eliminated by combining the results of the two squares. When this is done, the totals of the breakage rates for the four basic treatments are:

	Ingredient	
	A	*B*
1 *unit*	32·5	12·4
2 *units*	26·7	14·4

This is a 2 × 2 table of the kind analysed in section 10.3, and may be analysed into parts due to quantity, ingredient and the residual, each with one degree of freedom. Here the residual is the interaction between quantity and ingredient. For inclusion in Table 10.9 and comparison with the sums of squares given there, the sums of squares of this subsidiary analysis must be divided by 8, the number of readings per total. The same results would be obtained if the analysis were performed on the treatment *means* and the sums of squares were multiplied by eight. This is equivalent to performing the summations over all individual observations, and follows the lines of previous analyses. The results are in rows (7), (8) and (9) of Table 10.9.

To obtain the quantity-yarn interactions, we may eliminate the ingredient effects and obtain the following totals:

	Yarn	
	X	*Y*
1 *unit*	33·7	20·2
2 *units*	25·7	15·4

The analysis of these totals with the sums of squares divided by eight gives variances due to: yarn [row (1) of Table 10.9], quantity [row (7)] and the quantity-yarn interaction [row (10)].

Similarly, the ingredient-yarn interaction entered in row (11) of Table 10.9 may be obtained from the following totals:

	Ingredient	
	A	*B*
Yarn X	39·1	20·3
Yarn Y	20·1	15·5

The second-order interaction between the three factors is most easily obtained by subtracting the treatment effects so far determined from the total effect in row (4). This is given in row (12).

The variances in rows (7) to (12) of Table 10.9 are due to the various treatment effects plus the experimental error, so that they may all be tested against the variance in row (5) for significance. The main effects of quantity and ingredient, and the interaction between ingredient and yarn are significant, and we may express the final results by the following mean values of the breakage rates per 10 000 picks:

1 unit	..	..	3·4,	*2 units*	..	..	2·6,
Ingredient A	*yarn X*	..	4·9,	*Ingredient B*	*yarn X*	..	2·5,
	yarn Y	..	2·5,		*yarn Y*	..	1·9.

For comparing the quantities, since there are 16 readings per mean, the standard error of the difference is $\pm\sqrt{2\times 0{\cdot}712/16} = \pm 0{\cdot}30$, and the corresponding standard error of a difference between ingredients on one yarn is $\pm\sqrt{2\times 0{\cdot}712/8} = \pm 0{\cdot}42$. The difference between ingredients on yarn Y is not significant.

Thus, the final conclusions from these experiments are that the use of two units of either ingredient instead of one reduces the warp breakage rate slightly on both yarns, that on yarn X ingredient A gives a considerably lower breakage rate than B and that on yarn Y there is little or no difference in response to the two ingredients. It is interesting to note that the breakage rate is higher on yarn X than Y, and although we have not tested to see whether this is due to the yarns or to loom and period variations, it may be that the higher breakage rate is necessary to give sensitivity to the form of ingredient.

The computations for this example may be readily performed by the method of section 10.3, squaring and summing the various totals and dividing by the numbers of original observations in the totals. Before combining the results of the two Latin squares, it is well to perform the analyses separately to see if the two residual variances differ significantly. If they do, the tables should not be combined. We have ascertained that the two residuals are substantially the same, but have not given the results.

When possible, it is advisable to have more variations of each factor than were used in this experiment. With so few degrees of

freedom only marked effects are significant. More factors may be superimposed and higher order interactions investigated. Fortunately interactions of a high order are seldom significant; if they were often of importance experimental investigations would be very cumbersome and generalisations would be difficult.

Statistically, an interaction between factors appears as a residual in an analysis, and in the same way any residual may be regarded as an interaction. For example the residual term in Table 10.9 arising from the first analysis of Table 10.8 is in fact due to experimental errors and interactions between treatments, looms and periods. If a trial were made with uniformly sized warps, the residual would be less than that of Table 10.9 by the variance due to these interactions. Where such interactions exist, the assumptions underlying the analysis of a Latin square are not satisfied and it is advisable to arrange the treatments in random groups.

General Discussion

10.6. Technical literature, particularly that relating to agriculture, horticulture, forestry, animal breeding and other branches of applied biology, abounds in examples of the application of the foregoing arrangements and developments of them to practical experiments. A fairly detailed account of the subject treated from the points of view of the agriculturist is given by Sanders and Wishart (1935). Practically the whole of this experimental technique owes its inspiration to Professor R. A. Fisher, and readers are strongly recommended to refer to his book *The Design of Experiments* (1936*b*). There, Professor Fisher deals with the fundamental principles of experimental procedure and treats of various methods in detail.

Of the possible methods of arrangement, the most suitable will depend on the particular field of inquiry, on technical considerations and on the nature and extent of the variations in the untreated subject of experiment. In this connection, statistical surveys or uniformity trials are very valuable.

Results obtained from any particular experiment are subject to the obvious limitation that they do not necessarily apply under conditions other than those obtaining for the trial. For example, in a cereal variety trial, variety A may be significantly better than variety B, but it does not follow that A will always be better than B, e.g. in different localities and weathers, nor even that A will be better than B on the average. If it is desired to know whether A

gives a better yield than B on the average for all years and for several localities, the trial must be repeated in those localities and for several years, and the mean difference between the varieties be compared with a standard error obtained, not from a mean residual variance within the trials but from a residual variance *between* the trials. This point needs emphasis, because it may happen in practice that a highly significant result based on a single trial has only a narrow range of application.

This is one reason why multiple-factor experiments such as that exemplified in section 10.5 are advocated. The average effect of increasing the amount of the ingredients investigated there is established not only for one ingredient, but for two ingredients and yarns, and the applicability of the results is correspondingly widened.

In all the arrangements we have considered in this chapter, the various factors have balanced so that as far as possible each individual of any one factor is associated once with each individual of another. This has made the analysis much easier than it would otherwise have been. Equations (10.5) and (10.8) were much simplified because certain product terms were zero, so that the deviations due to each factor could easily be separated and estimated independently. This property is termed *orthogonality* and is highly desirable.

Sometimes, however, accidents occur and individual readings are lost. In many such instances the missing readings may be estimated by the method of least squares, so that the analysis of variance is still possible. In some experiments, too, a complete arrangement involving every combination of all factors may be unwieldy and unnecessary. Then certain combinations may be omitted, and the corresponding treatments are said to be *confounded*. For example, in the arrangement of Table 10.8, the yarn effect is confounded with the loom and period effects, and there is no possibility of measuring the interactions. This aspect is dealt with by Fisher (1936*b*), and the treatment of a number of incomplete and confounded arrangements is worked out in a number of papers by Yates (1933 and 1936).

Whether the experimental arrangement is balanced or unbalanced, it should be carefully designed, otherwise serious inefficiency may result or two important effects may be confounded and so be unmeasurable. The time to consult statistical principles is before

the experiment is planned, not after the results are obtained and are in confusion.

It is well to emphasise again that all the tests of significance and estimates of corrected variance in this chapter are based on the assumption of a homogeneous residual variation; i.e. it is assumed that the residual variance is, within the limits of random errors, the same for all blocks, rows, columns and factors that are combined.

CHAPTER XI

MULTIPLE AND PARTIAL CORRELATION AND REGRESSION

Partial Correlation

11.1. We showed in the last chapter how the heterogeneity of variability has an important effect on the theory of errors, and in this chapter shall study its effect on correlation.

Table 11.1 shows the mean grain and straw yields for each of

TABLE 11.1

GRAIN AND STRAW YIELDS

Treatment	Block 1		Block 2		Block 3		Block 4	
	Grain	Straw	Grain	Straw	Grain	Straw	Grain	Straw
A	620	242	646	321	681	261	644	317
B	644	267	745	382	542	201	711	316
C	523	215	713	330	686	298	688	381
D	601	212	693	292	685	265	714	255
E	664	322	693	370	666	284	516	323
F	514	200	637	261	697	259	710	361
G	550	260	708	318	663	266	673	340
H	521	203	661	275	594	207	730	331
	Block 5		**Block 6**		**Block 7**		**Block 8**	
A	706	255	615	331	552	216	726	295
B	705	280	637	285	543	200	646	309
C	692	300	612	294	635	256	748	284
D	699	238	697	309	701	283	746	324
E	656	232	663	393	657	351	683	363
F	633	234	595	258	697	306	712	376
G	671	362	626	400	655	276	671	385
H	625	229	644	266	745	276	747	328

64 plots (the units do not concern us here), there being eight different manurial treatments and eight replicates on each. The data are selected from a paper by Eden and Fisher (1927), and are merely used here for illustrative purposes; readers who are interested in the agricultural aspect must consult the original paper. Further, we

shall for the moment ignore the fact that the sample is small with comparatively large sampling errors, and shall treat the statistical constants as though they were almost exact. The treatments were distributed at random within blocks and we have in Table 11.2 analysed the variance of grain and straw yields into three parts; the fifth, eighth and ninth columns should be ignored for the moment. The block variation is much greater than the residual for both grain and straw, and so is the treatment variation for straw, while the variance of grain yields is less for the treatments than for the residual (although not significantly so); thus the variations in the yields of Table 11.1 are heterogeneous. Now in order to find the relation between grain and straw yields, we may correlate the sixty-four pairs of readings, and the crude correlation of the actual readings gives a coefficient of $+0{\cdot}524$, indicating a positive relationship. This, however, is not the whole story; the variations are produced by three groups of causes, changes in soil fertility, changes in treatments and that complex of unknown causes which we call random, and it is unlikely that the relationship between grain and straw yields is the same for the three types of variation. Indeed, in a general way, plots that produce most grain might be expected to produce most straw; but the treatments which had an effect on the straw yield had none on the grain, and this relationship cannot hold for treatment variations. It is thus necessary to separate out the effects and to find several correlations, and this is done by a fairly straightforward extension of the analysis of variance.

Let us suppose first that the block and residual variations are due to the same thing (soil variability), and that we wish to separate them from those due to manurial treatments, and to correlate them. This may be done quite straightforwardly by finding the 64 pairs of deviations from the treatment means and correlating them, by finding the sums of their products and squares. We may also correlate the eight pairs of treatment means (or totals). Using the same notation as before, but calling the grain yield y and the straw yield x, the values of an individual pair of readings in the sth treatment as deviations from the grand means are

$$\left.\begin{aligned}(x-\bar{x}) &= (x-\bar{x}_s)+(\bar{x}_s-\bar{x})\\ \text{and}\qquad (y-\bar{y}) &= (y-\bar{y}_s)+(\bar{y}_s-\bar{y}),\end{aligned}\right\} \quad . \quad . \quad . \quad (11.1)$$

TABLE 11.2.—ANALYSIS OF VARIANCE AND CO-VARIANCE OF GRAIN AND STRAW YIELDS

Source of Variations	Degrees of Freedom	Sums of Squares		Sums of Products	Variances		Co-variance	Correlation Coefficient
		Grain	Straw		Grain	Straw		
Blocks	7	86 045·8	75 841·5	56 073·6	12 292·3	10 834·5	8 010·5	+0·694
Treatments ..	7	12 496·8	32 985·0	−6 786·6	1 785·3	4 712·1	−969·5	−0·334
Residual ..	49	136 972·6	71 496·1	58 549·0	2 795·4	1 459·1	1 194·9	+0·592
Total ..	63	235 515·2	180 322·6	107 836·0	—	—	—	+0·524

TABLE 11.3.—ANALYSIS OF VARIANCE AND CO-VARIANCE

Maximum and Minimum Temperatures

Source of Variations	Degrees of Freedom	Sums of Squares		Sums of Products	Variances		Correlation Coefficient
		Maximum	Minimum		Maximum	Minimum	
Within years ..	1 469	38 360 589	28 565 319	8 483 383	26 113	19 445	+0·256
Between years ..	48	14 127 693	3 549 394	3 772 672	294 327	73 946	+0·533
Total ..	1 517	52 488 282	32 114 713	12 256 055	—	—	+0·298

and their product summed for the sth treatment is

$$S'(x-\bar{x})(y-\bar{y}) = S'(x-\bar{x}_s)(y-\bar{y}_s) + n_s(\bar{x}_s-\bar{x})(\bar{y}_s-\bar{y}) + (\bar{y}_s-\bar{y})S'(x-\bar{x}_s) + (\bar{x}_s-\bar{x})S'(y-\bar{y}_s),$$

where n_s is the number of plots per treatment. The last two terms are zero since $S'(x-\bar{x}_s)$ and $S'(y-\bar{y}_s)$ are the sums of deviations from the mean and are zero, so this equation, when summed further over all treatments, gives

$$S(x-\bar{x})(y-\bar{y}) = \underset{s}{S}S'(x-\bar{x}_s)(y-\bar{y}_s) + \underset{s}{S}\, n_s(\bar{x}_s-\bar{x})(\bar{y}_s-\bar{y}). \quad . \quad . \quad (11.11)$$

The term on the left-hand side is the sum of products used in correlating the crude deviations from the grand mean, the first term on the right-hand side is used in correlating the deviations from the treatment means and the second term in correlating the treatment means themselves. This equation is exactly parallel to equation (6.6) on p. 132 for the analysis of variance, and may similarly be entered in a table of analysis of *co-variance*. The degrees of freedom are reckoned up in the same way, and the sum of products divided by the number of degrees of freedom gives the mean product or co-variance. The sums of squares and variances can be entered in the same table, and the co-variance of any cause of variation when divided by the square root of the product of the variances [equation (7.4), p. 165] gives the corresponding correlation coefficient.

We can deal with the co-variance in exactly the same way as the variance, using all the arguments and equations previously used, except that co-variance may sometimes be negative while variance must always be positive. We saw in the last chapter that variance may be analysed into more than two parts, and now will use the same methods to analyse the co-variance of the yields of Table 11.1 into three parts. For a single plot in the sth treatment and tth block,

$$(x-\bar{x}) = (\bar{x}_s-\bar{x}) + (\bar{x}_t-\bar{x}) + x',$$
$$(y-\bar{y}) = (\bar{y}_s-\bar{y}) + (\bar{y}_t-\bar{y}) + y',$$

where x' and y' are residual deviations; and by a similar argument to that used above, summing over all treatments and blocks,

$$S(x-\bar{x})(y-\bar{y}) = \underset{s}{S}n_s(\bar{x}_s-\bar{x})(\bar{y}_s-\bar{y}) + \underset{t}{S}n_t(\bar{x}_t-\bar{x})(\bar{y}_t-\bar{y}) + Sx'y', \quad . \quad . \quad (11.2)$$

where n_s is the number of plots in a treatment and n_t the number in a block. This is exactly parallel to equation (10.5), p. 215,* and the terms have been entered in Table 11.2 under the *sums of products* column, and when divided by the degrees of freedom, as co-variances. For convenience of computing, equation (7.61), p. 166, may be applied to find the terms of the above; i.e.

$$S(x - \bar{x})(y - \bar{y}) = Sxy - N\bar{x}\bar{y},$$
$$\underset{s}{S} n_s(\bar{x}_s - \bar{x})(\bar{y}_s - \bar{y}) = \underset{s}{S} n_s\bar{x}_s\bar{y}_s - N\bar{x}\bar{y},$$

and so on; but particular care must be taken of signs, as co-variance may be negative. The correlation coefficients in Table 11.2 have been found from the sums of squares and products, and whereas the coefficient for block and residual variations is about + 0·6, for the treatment variations it is — 0·3; the difference is important. The residual deviations are independent of block and treatment differences, and their correlation coefficient is a *partial* coefficient with both block and treatment factors eliminated. The eight treatment totals are independent of the blocks and the block totals are independent of the treatments, but both are to some extent affected by the residual deviations, and although their correlation coefficients are in some degree partial, only *one* factor has been eliminated. However, as there are several plots per block and treatment, the effects of these two factors predominate over the residual in the correlation coefficients.

The maximum and minimum temperatures in Table 7.4, p. 147, are daily readings taken in the month of August for 49 years, and we may now be led to inquire if the variations between and within years are of the same nature. Fisher and Hoblyn give the sums of squares and products, and we repeat them in Table 11.3, together with the correlation coefficients.† The variances show that the between-year variations are real, and the correlation coefficients show that the association between maximum and minimum temperatures is stronger for these variations than for those within the year; that is to say, if we know the average maximum temperature for the month of August in any one year we can estimate the average minimum for that month with a little greater accuracy than we can the minimum for any one day, knowing the corresponding maximum.

* If n_s and n_t are constant, they may be placed outside the summation sign.

† The data of Table 11.3 have been calculated from Fisher and Hoblyn's full correlation table and not from the condensed Table 7.4.

is given in the last row of Table 11.21. The variance due to treatments is now significantly greater than the residual, so that after correction for the straw yield, the treatments have a significant effect on grain yield. This is partly because of the considerable reduction in the residual variance resulting from the important partial correlation between the two yields. We make no attempt to give a biological interpretation of this result.

If a is the regression of grain on straw yield for the residual

TABLE 11.21

ANALYSIS OF VARIANCE OF GRAIN YIELD AFTER CORRECTION FOR REGRESSION ON STRAW YIELD

Source of Variations	Sum of Squares	Degrees of Freedom	Variance
Total within blocks ..	123 825·1	55	2 251·4
Residual	89 026·1	48	1 854·7
Treatments (difference) ..	34 799·0	7	4 971·3

deviations as estimated from the third row of Table 11.2, the treatment means corrected for variations in straw yield are

$$\bar{y}_s \text{ (corrected)} = \bar{y}_s - a(\bar{x}_s - \bar{x}).$$

Here,

$$a = \frac{Sx'y'}{Sx'^2} = 0{\cdot}818\,91.$$

If a were the true population value, the standard error of the difference between any two corrected means would be calculable from the corrected residual variance in Table 11.21. As it is, the error in a also contributes somewhat to the standard errors of the corrected treatment means. It is for this reason that the variance due to treatments in Table 11.21 cannot be obtained directly from the corrected means.

A discussion of the practical aspects of this application of the analysis of co-variance technique to agricultural experiments is given by Fisher (1936), Sanders and Wishart (1935).

Elimination of a Linear Factor

11.2. In section 11.1, when finding the partial correlation between grain and straw yields with block and treatment effects eliminated, we measured the yields as deviations from discrete block and treatment means and correlated the deviations. Sometimes, however, the factor to be eliminated can be described quantitatively and its relation to the variates to be correlated can be expressed by regression lines instead of by discrete means. Then the deviations from these lines may be correlated. We shall deal with partial correlation when the regression lines are straight.

If the quantities are x, y and z and z is to be eliminated, let the values of x and y lying on the regression lines relating them to z be X_z and Y_z. Then equation (11.11) becomes

$$S(x - \bar{x})(y - \bar{y}) = S(x - X_z)(y - Y_z) + S(X_z - \bar{x})(Y_z - \bar{y}).$$

The deviations from the regression lines may be found explicitly, and from them the partial correlation coefficient, or if the regressions are linear, the following formula may be used,

$$r_{xy\,.\,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \quad . \quad . \quad . \quad (11.3)$$

where $r_{xy\,.\,z}$ is partial correlation coefficient between x and y with z eliminated, and r_{xy}, r_{xz} and r_{yz} are the three total correlation coefficients between x, y and z. This equation is proved in the appendix to this chapter. By interchanging the suffixes, the partials between x and z and between y and z can be found from the same formula. In using equation (11.3), tables by Miner (1922) of $\sqrt{1 - r^2}$ for different values of r will be found very useful.

Mumford and Young (1923) give the correlation coefficients in Table 11.4, based on measurements taken on 1 110 boys; we will assume the regressions to be linear. It would appear at first sight that there is a strong tendency for tall boys to have a large vital capacity (the correlation coefficient is 0·835), but the older boys tend to be taller and (as might be expected) to have a larger vital capacity. It is reasonable, therefore, to suppose that a good deal of the apparent association between vital capacity and height may be due to the fact that the taller boys are also older, and before the true connection can be found, correction must be made for age. Similar arguments apply to the correlations between vital capacity and weight and between height

and weight, and, using equation (11.3) to find the three partial correlations with age eliminated, we obtain the results in the upper part of Table 11.41. The first three coefficients show that the relationships between vital capacity, height and weight continue to exist, even though the age is kept constant, although they are a little less

TABLE 11.4*

CORRELATION COEFFICIENTS

Vital capacity and height	+0·835	Vital capacity and weight	+0·851
Vital capacity and age..	+0·662	Weight and age ..	+0·701
Height and age ..	+0·714	Height and weight ..	+0·897

* We only include a portion of the data in the paper.

strong. We are still, however, far from a complete knowledge; part of the association between vital capacity and height may really be an indirect consequence of the fact that tall boys are also heavy, and weight may be the determining factor, or, conversely, height may have the direct and weight the indirect connection with vital capacity. Again we may use the formula of equation (11.3) to eliminate further, and in turn, weight and height, and the results are in the lower part of Table 11.41. The correlations are now much reduced, but are

TABLE 11.41

PARTIAL CORRELATION COEFFICIENTS

Vital capacity and height (age eliminated)	.. ..	+0·690
Vital capacity and weight (age eliminated)	.. ..	+0·724
Height and weight (age eliminated)	.. .. ..	+0·794
Vital capacity and height (weight and age eliminated)	..	+0·271
Vital capacity and weight (height and age eliminated)	..	+0·399

still real, and it appears that the association of vital capacity with height is a little less important than that with weight. Thus, by means of equation (11.3), any number of factors can be eliminated one at a time, and ultimate partial correlation coefficients can be found. In using the formula, it is advisable to retain more significant figures than will be required in the final answer, as errors due to too drastic approximation tend to multiply themselves. Mumford

and Young in their paper consider other factors (stem length, chest girth), and the conclusions even of the second part of Table 11.41 cannot be regarded as final.

Elimination of a Non-linear Factor

11.3. If the regression between any two factors differs much from linearity, equation (11.3) is not valid, although it may often be used as a fair approximation when the departure from linearity, although statistically significant, is not very great. It is always possible, of course, to fit curved regression lines of x and y on z, to find the actual deviations of x and y from those lines and to correlate the deviations, but in such circumstances it is obviously desirable (and *necessary*, if tests of significance are to be applied) that both x and y should be fitted in the same way and to the same degree. Any method of smoothing may be used to determine the regression curve, but some such form as the polynomial, which can be found by the method of least squares, is preferable so that the number of degrees of freedom absorbed in fitting may be known. If the polynomial is used (say) to the third degree in z, and x and y are the variables which are to be partially correlated, it is permissible to regard x, y, z, z^2 and z^3 as five variables related linearly (finding the individual values of z^2 and z^3 from those of z), to correlate them in every possible way, and then to eliminate z^3, z^2 and z in turn by repeated applications of equation (11.3) as shown in the last section. If there is only one value of the dependent variable to every one of that which is to be eliminated, and a polynomial can be fitted according to Fisher's system as outlined in section 9.31, the operation of correlating residuals is very easy; we deal with it in the next section.

11.31. We found constants A, B, C, etc., of the polynomial [equation (9.6), p. 197] and said that the sums of squares of the residuals were found by subtracting in turn from the crude sum of squares of the observations, quantities NA^2, etc. (section 9.31); similarly the sums of products of the residuals after fitting any number of terms of the polynomial are obtained by subtracting successively from the crude sum of products the quantities,

$$NA_1A_2,\quad \frac{N(N^2-1)}{12}B_1B_2,\quad \frac{N(N^2-1)(N^2-4)}{180}C_1C_2,$$
$$\frac{N(N^2-1)(N^2-4)(N^2-9)}{2\,800}D_1D_2,\ \text{etc.} \quad . \quad (11.4)$$

(paying due regard to sign) where A_1, B_1, etc., are the constants of the polynomial of the first factor on the eliminated one and A_2, B_2, etc., are the constants of the polynomial relating the second factor to the eliminated one. If a curve of zero order is fitted, i.e. if the deviations are measured from the mean, the sum of products of residuals is $S(xy) - NA_1A_2$, and this is the ordinary method of correcting to the mean by subtracting $N\bar{x}\bar{y}$; the subtraction of the other terms for curves of higher order may be regarded as equivalent to correcting to a moving mean. These product terms are just as much part of the analysis of co-variance as the square terms are of the analysis of variance, and we have inserted them in Table 11.5 for the two series of protein percentages, the original data being in Table 8.1, p. 174, and the constants being given on p. 199; Table 11.5 corresponds exactly to Table 9.4 (p. 200), and the two should be read in conjunction. It would be a good and convincing exercise if the reader were to determine the 58 protein contents given by the two cubic regressions, find the actual residuals, and see if their sums of squares and products agree with those in Tables 9.4 and 11.5. As before, the sum of products at any row may be used with the corresponding sums of squares to determine a partial correlation coefficient. We will now use these data to illustrate another type of statistical problem.

Spurious Correlation of Time Series

11.4. An investigator may ask if there is any common factor (perhaps weather) which affects the protein contents of both series of corn in the same way, and it is natural to use the correlation coefficient to examine this question. The crude correlation coefficient as found from the "totals" rows of Tables 9.4 and 11.5 is $-0{\cdot}638$, and would at first sight appear to indicate that if the common factor affects protein content, it acts on the two series in opposite ways. We have to remember, however, that because of the way in which seed was selected for the two series, one increased in protein content while the other decreased, and this predominating effect gives a negative correlation coefficient. In order to investigate the effect of the possible common factor, we must first eliminate the variation in protein with time (due to seed selection), and then correlate the residuals. First, we will assume as an approximation that the regressions on time are linear, with correlation coefficients of $+0{\cdot}862$

and $-0{\cdot}708$, and using equation (11.3) (time is z and the protein contents are x and y), find the partial correlation to be

$$\frac{-0{\cdot}638+0{\cdot}862\times0{\cdot}708}{\sqrt{(1-0{\cdot}862^2)\,(1-0{\cdot}708^2)}}=-0{\cdot}08,$$

which, on such a small sample, is not significant. We have seen, however, that the regression of the two protein contents on time departs considerably from linearity, and equation (11.3) is useful only as an approximation; but we may use the analyses of Tables 9.4

TABLE 11.5

ANALYSIS OF CO-VARIANCE OF PROTEIN CONTENT

Source of Variation	Degrees of Freedom	Sums of Products
First order regression ..	1	−33·01
Second order regression ..	1	− 2·86
Third order regression ..	1	− 2·18
Residual	25	+ 3·52
Total	28	−34·53

and 11.5 to correlate the residuals after fitting cubic equations. This correlation coefficient is

$$\frac{+3{\cdot}52}{\sqrt{12{\cdot}34\times12{\cdot}06}}=+0{\cdot}29,$$

and although still too small to be significant on such a small sample, does suggest that some common factor varying from year to year may have been affecting the annual residual fluctuations of protein yield in the two series in the same way.*

Such problems, in which it is desired to correlate two series of observations extended in time or space, have received a good deal of attention from statisticians; the example in section 7.3 of the correla-

* We have treated the first observation in each of the two series as independent, but they are actually the same protein content of a common crop. This does not affect the constants much, nor, in this instance, our conclusions.

tion between age at death and the proportion of marriages contracted in the Church of England comes into this category. Essentially the problem is a special case of partial correlation, with time (or space) as the variable to be eliminated, and the discussion has centred round the development of suitable means of expressing the regression and obtaining residuals; there is most difficulty in economic statistics where factors are apt to vary periodically (or in waves), but for most biological data extending over only a few years, the method we have outlined is adequate. Even if no such obvious effect as that in the example just worked is present, the possibility of some unsuspected time-effect in data collected over some period of time should always be considered.

Partial and Multiple Regression

11.5. With every partial correlation coefficient there goes a *partial regression coefficient*, which is obtained by dividing the sum of products by the appropriate sum of squares of the independent variable and which expresses the amount of change in one factor associated with a second when the other factors are constant. The regression of grain on straw yield (Table 11.2) for the residual variations is the value of a on p. 250, viz.

$$\frac{58\,549{\cdot}0}{71\,496{\cdot}1} = 0{\cdot}82$$

unit of grain associated with a change of 1 unit of straw per plot.

When there are several quantities all linearly related, partial regressions are best found by fitting a multiple regression equation. If y is the dependent variable and x_1, x_2, x_3 . . . are the independent variables, the linear regression of y on the x's is expressed by an equation of the form,

$$Y' = ax_1' + bx_2' + cx_3' + \ldots^{*} \quad . \quad . \quad . \quad (11.5)$$

Equation (11.5) is called the *multiple regression* formula of y on x_1, x_2, x_3, etc., and the constants a, b, c, etc., are the partial regression coefficients, showing how much unit changes in the individual x variables affect y, independently and directly. The constants, a, b, c,

* x' and y' are deviations from the grand means.

etc., may be found by the method of least squares, and this leads to a series of linear equations,

$$\left.\begin{aligned} Sy'x'_1 &= aSx'^2_1 + bSx'_1x'_2 + cSx'_1x'_3 + \ldots \\ Sy'x'_2 &= aSx'_1x'_2 + bSx'^2_2 + cSx'_2x'_3 + \ldots \\ Sy'x'_3 &= aSx'_1x'_3 + bSx'_2x'_3 + cSx'^2_3 + \ldots \\ \cdot \quad & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned}\right\} \quad . \quad (11.6)$$

The summations have to be found, of course, from the data, and the equations have to be solved for a, b, c, etc. These are not unlike equations (9.4) given on p. 195 for determining the constants of a curved regression line; and indeed the two processes are somewhat the same, for in the earlier one we are finding the multiple regression of y on x, x^2, x^3, etc.

We shall illustrate this by finding from the data of Table 11.6 (Harris, 1931) a formula for predicting the loaf volume from the protein content of the flour and the percentage of it extracted by potassium bromide solution. The forty-four wheats were analysed and subjected to the baking test to give the data of the table. If we call the loaf volume y, and the protein percentages x_1 and x_2, the sums of squares and products are as given in the following equations,

$$\begin{aligned} 1\,462{\cdot}33 &= \quad 107{\cdot}55\,a - 161{\cdot}64\,b \\ -\,2\,022{\cdot}85 &= -\,161{\cdot}64\,a + 473{\cdot}19\,b. \end{aligned}$$

These, solved for a and b, lead to the regression equation,

loaf volume in c.c. $= 14{\cdot}738\,5 \times$ *crude protein per cent.*
$+ 0{\cdot}759\,7 \times$ *protein extracted by KBr per cent.*

all quantities being measured as deviations from their means. If we were to use the crude protein percentage only, a would then be the ordinary linear regression, and equal to

$$\frac{1\,462{\cdot}33}{107{\cdot}55} = 13{\cdot}6 \text{ c.c.},$$

for a change of 1 per cent. in protein.

The Multiple Correlation Coefficient

11.6. Just as we need the correlation coefficient to give an idea of the accuracy of a linear regression formula with a single term, so we need a constant to specify the accuracy of a multiple regression formula, and this is provided by the *multiple correlation coefficient*.

This may be found by correlating the predicted values (Y) given by the formula with the actual ones (y); or alternatively, the sum of squares of the deviations of y from the regression is $(1 - R^2)$ multiplied by the sum of squares of the deviations of y from the grand mean, where R is the multiple correlation coefficient. Thus, multiple correlation is another case of the analysis of variance, analysing

TABLE 11.6

Loaf Volume (c.c.)	Crude Protein, Per cent.	Protein Extracted by KBr, per cent.*	Loaf Volume (c.c.)	Crude Protein, Per cent.	Protein Extracted by KBr, per cent.
540	12·0	31·2	472	12·6	21·8
475	11·5	31·4	500	10·4	31·6
450	11·0	31·3	520	13·8	28·8
455	12·6	27·9	525	13·6	26·7
465	11·9	31 4	490	11·5	30·8
470	12·0	30·2	505	10 6	31·5
440	11·7	31·9	475	10·7	31·5
490	12·1	31·8	545	14·2	28·8
475	12·9	31·0	534	13·4	27·4
465	11·3	31·7	490	11·0	29·3
520	14·7	28·3	507	10·8	29·4
525	15·2	20·5	530	12·8	28·2
450	10·5	28·7	505	12·2	29·7
500	13·1	27·7	495	12·5	28·9
445	10·1	31·8	515	12·4	30·2
443	11·2	28·9	505	14·7	21·4
463	11·3	25·5	535	14·5	22·7
446	11·5	30·0	445	9·0	34·9
500	12·9	28·5	470	11·2	29·7
487	10·4	25·5	435	9·1	34·0
502	12·7	25·5	475	10·8	29·2
529	16·0	21·8	505	13·9	26·0

* This is the percentage of the crude protein extracted by KBr.

that of y into two parts, one associated with the regression and the other a residual, and the degrees of freedom of the former are equal to the number of terms in the regression equation. We have entered these terms in Table 11.7, assuming the size of sample to be N and the number of terms [x's in equation (11.5)] to be m'. If there is only one term on the right-hand side of the regression equation, $m' = 1$, and Table 11.7 reduces to the form for the correlation of

TABLE 11.7

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Multiple Regression	$S(Y-\bar{y})^2 = R^2 S(y-\bar{y})^2$	m'	$\frac{R^2}{m'} S(y-\bar{y})^2$
Residual	$S(y-Y)^2 = (1-R^2) S(y-\bar{y})^2$	$N-m'-1$	$\frac{1-R^2}{N-m'-1} S(y-\bar{y})^2$
Total	$S(y-\bar{y})^2$	$N-1$	—

TABLE 11.71

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom
Regression with m_2' constants } Regression with m_1' constants	$R_2^2 S(y-\bar{y})^2$	m_2'
Difference between regressions } Regression with m_1' constants	$(R_1^2 - R_2^2) S(y-\bar{y})^2$ } $R_1^2 S(y-\bar{y})^2$	$m_1' - m_2'$ } m_1'
Residual from regression with m_1' constants ..	$(1-R_1^2) S(y-\bar{y})^2$	$N-m_1'-1$
Total	$S(y-\bar{y})^2$	$N-1$

two variables (Table 7.6, p. 158), where $R = r$. On the other hand, if we compare Tables 11.7 and 9.2 (p. 193) we see that they are of essentially the same form, R^2 corresponding to η^2, and m' to $(m - 1)$, N being the same; thus, the multiple correlation coefficient and the correlation ratio are essentially analogous constants, both measuring association, the former between one variate and a number of others as expressed by a multiple linear regression formula, and the latter between one variate and another expressed as a series of array means or as a curved regression line. We have already pointed out the analogy between the multiple regression formula and the curved regression line, and now see that it is complete. Like the correlation ratio, the multiple correlation coefficient is essentially positive, and becomes larger as the number of degrees of freedom associated with it (m') increases; indeed, if there were as many constants as there are independent deviations of y in the sample ($m' = N - 1$), R would equal unity. Hence, all that we have written about the correlation ratio is true of the multiple correlation coefficient.

In order to find R, it is not necessary to find the individual regression values of y, but the following relation may be used

$$SY'^2 = aSy'x'_1 + bSy'x'_2 + cSy'x'_3 + \ldots * \quad . \quad (11.7)$$

For our wheat quality and loaf volume data,

$$SY'^2 = R^2Sy'^2 = 14{\cdot}738\,5 \times 1\,462{\cdot}33 - 0{\cdot}759\,7 \times 2\,022{\cdot}85 = 20\,015{\cdot}8,$$

* Squaring equation (11.5),

$$\begin{aligned} Y'^2 = {} & ax'_1[ax'_1 + bx'_2 + cx'_3 + \ldots] \\ & + bx'_2[ax'_1 + bx'_2 + cx'_3 + \ldots] \\ & + cx'_3[ax'_1 + bx'_2 + cx'_3 + \ldots] \\ & + \ldots\ldots\ldots\ldots \\ = {} & a[ax'^2_1 + bx'_1x'_2 + cx'_1x'_3 + \ldots] \\ & + b[ax'_1x'_2 + bx'^2_2 + cx'_2x'_3 + \ldots] \\ & + c[ax'_1x'_3 + bx'_2x'_3 + cx'^2_3 + \ldots] \\ & + \ldots\ldots\ldots\ldots \end{aligned}$$

and summing,

$$\begin{aligned} SY'^2 = {} & a[aSx'^2_1 + bSx'_1x'_2 + cSx'_1x'_3 + \ldots] \\ & + b[aSx'_1x'_2 + bSx'^2_2 + cSx'_2x'_3 + \ldots] \\ & + c[aSx'_1x'_3 + bSx'_2x'_3 + cSx'^2_3 + \ldots] \\ & + \ldots\ldots\ldots\ldots \end{aligned}$$

Equation (11.7) follows from this when the values of equation (11.6) are substituted for the terms in the brackets.

and since

$$Sy'^2 = 41\,246{\cdot}8, \quad R^2 = 0{\cdot}485\,3 \quad \text{and} \quad R = 0{\cdot}697.$$

Correlating the loaf volume with the crude protein content only, we find $r = +\,0{\cdot}694$, and thus see that although it absorbs one more degree of freedom, the multiple regression equation is scarcely better than that using protein content as a single factor. We deal with tests for the significance of such a difference in section 11.72.

Errors of Sampling in Multiple and Partial Correlation

Partial Correlation Coefficient

11.71. Fisher (1924*c*) has shown that the distributions of the partial and the total correlation coefficients are the same, when they are estimated from deviations having the same number of degrees of freedom. Thus, if there are N pairs of observations and m factors have been eliminated (leaving $N - m - 1$ degrees of freedom), the partial correlation coefficient is distributed like that of the total coefficient based on $(N - m)$ pairs of observations.

The residual correlation coefficient between grain and straw yield in Table 11.2 is based on 49 degrees of freedom and may be treated like an ordinary correlation based on 50 pairs of observations, while the correlation for treatment variations is based on 7 degrees of freedom, and may be regarded as arising from a sample of 8; we will use Fisher's z' transformation (section 8.2) to test the significance of the difference between them. The values of z' are 0·681 and − 0·347 with a difference of 1·028; the standard error of this is

$$\pm\sqrt{\tfrac{1}{5} + \tfrac{1}{47}} = \pm\,0{\cdot}47,$$

and the difference is significant.

We found in section 11.4 that the partial correlation coefficient between the residuals of protein contents of two series of corn after the time trend had been eliminated by a cubic was + 0·29, and since there were 29 observations with four terms eliminated (including the mean) we must test this as though it were based on 26 pairs of observations. Using Fisher's table of values of the correlation coefficient at different levels, we should enter it at $n = 24$; there is no value at this point, but we see that for $n = 25$, $r = 0{\cdot}32$ lies on the 0·1 level, and our observed value, being less even than this, is not significant.

For the same reasons as advanced in section 10.1 in connection with the mean, the sampling error of the correlation coefficient is only given by the usual formulæ when the correlation is between homogeneous and independent deviations arrived at after a complete analysis of co-variance.

Multiple Correlation Coefficient

11.72. In order to test the significance of a multiple correlation coefficient we may use the analogy with non-linear regression and employ the same test. From the analysis of variance of Table 11.7 we find

$$z = \log_e \sqrt{\frac{R^2(N - m' - 1)}{(1 - R^2)m'}}$$

and enter Fisher's tables at $n_1 = m'$ and $n_2 = (N - m' - 1)$. Wishart (1928) gives a table of values of the multiple correlation coefficient lying on the 0·05 and 0·01 levels of significance.

For the data of Table 11.6 (loaf volumes and wheat qualities), $R^2 = 0{\cdot}485$, $m' = 2$, $N = 44$,

$$z = \tfrac{1}{2} \log_e \frac{0{\cdot}485}{0{\cdot}515} \times \frac{41}{2} = 1{\cdot}48,$$

and this is even greater than the value of z lying on the 1 per cent. level when $n_1 = 2$ and $n_2 = 30$ (i.e. greater than 0·84). Hence this multiple correlation coefficient is significantly greater than zero.

We may also wish to test if a regression formula with a large number of constants gives a significantly closer prediction than one using fewer, and this may be done in exactly the same way as in section 9.2, where we tested for non-linearity of regression. If there are two regression formulæ involving m_1' and m_2' constants, m_1' being greater than m_2', and the multiple correlation coefficients are R_1 and R_2, the complete analysis of variance is as shown in Table 11.71, and since all the sums of squares are positive, R_1 must always be greater than R_2. If, however, the difference between R_1 and R_2 is the result of a real effect, the variance for the difference between regressions will be significantly greater than the residual, and using Fisher's test, we see if

$$z = \log_e \sqrt{\frac{(R_1^2 - R_2^2)\ \ (N - m_1' - 1)}{(1 - R_1^2)\ (m_1' - m_2')}}$$

is significant for degrees of freedom,

$$n_1 = m_1' - m_2' \quad \text{and} \quad n_2 = N - m_1' - 1.$$

Applying this to the data of Table 11.6, r^2 for protein content alone $= 0{\cdot}482$ $(m_2' = 1)$, while for the multiple regression, $R^2 = 0{\cdot}485$ $(m_1' = 2)$; hence

$$z = \log_e \sqrt{\frac{0{\cdot}003 \times 41}{0{\cdot}515}} = -0{\cdot}716,$$

and certainly is not significant. Thus, the extraction of some of the protein by potassium bromide has added nothing to the information provided by the crude protein content alone, as far as the prediction of loaf volume for this series of flours is concerned.

11.73. We will illustrate the further application of these tests by another example, which presents several features; the data are in Table 11.8. Fifty groups of varying numbers of cotton hairs were weighed and measured, and the weights of the groups were expressed as multiples of 10^{-8} grammes per centimetre. These were then swollen in caustic soda and placed in three classes, A, B and C, according to their appearance under the microscope. The problem is to determine if the three classes show real differences in average hair-weight per centimetre. A regression formula of the form of equation (11.5) may be obtained to express the weight of a group of hairs in terms of the number in each of the three classes; if we put y equal to the group weight, and x_1, x_2 and x_3 the numbers in the three classes, a, b and c are the corresponding mean hair-weights per centimetre. Similarly a second regression may be obtained in which only the average hair-weight for all classes, and the total number of hairs in the group are used; this is the ordinary case with one constant. Now the former regression, absorbing three degrees of freedom, will necessarily have associated with it a little more of the variance in weights of the groups than the second, but if the three groups really have different-hair weights, the equation with three constants will fit the data very much better than that with only one, and the difference in the associated variances will be *significantly* greater than zero. We may therefore investigate the problem by finding the regressions and variances, and testing the significance of the difference in the latter.

TABLE 11.8.*—NUMBERS OF COTTON HAIRS AND WEIGHTS (IN 10^{-8} GRAMMES)

Numbers of Hairs				Weight	Numbers of Hairs				Weight
A	B	C	Total		A	B	C	Total	
15·4	7·6	—	23	3 243	14·5	3·4	1·1	19	3 401
25·0	4·0	1·0	30	5 580	10·5	3·5	—	14	2 016
22·3	6·7	—	29	4 872	24·0	10·0	—	34	6 596
45·6	7·4	1·0	54	11 880	25·5	7·4	1·1	34	6 290
51·0	4·0	—	55	13 420	66·7	13·2	1·1	81	15 147
63·3	12·7	1·0	77	16 632	81·9	13·1	4·0	91	20 790
88·5	17·5	—	106	22 684	77·9	5·1	—	83	18 260
71·9	8·1	—	80	18 240	63·1	14·8	1·1	79	15 642
53·8	6·2	—	60	13 380	42·9	7·1	—	50	8 950
41·0	4·0	—	45	9 180	24·9	2·1	—	27	4 887
13·0	—	—	13	3 185	3·0	—	—	3	561
7·9	1·1	—	9	1 818	4·0	—	—	4	660
2·7	4·0	1·3	8	816	12·3	3·7	—	16	2 272
8·2	1·8	—	10	2 190	9·8	2·8	1·4	14	2 072
15·0	8·0	—	23	3 864	15·0	7·0	—	22	3 762
14·5	5·5	—	20	4 400	24·0	6·0	2·0	32	6 144
17·6	12·4	1·0	31	6 045	31·0	7·0	—	38	8 018
47·6	12·4	—	60	14 880	54·4	8·6	—	63	14 238
70·9	19·0	1·1	91	20 657	97·3	17·6	1·1	116	23 896
50·0	17·0	—	67	13 601	95·6	14·4	—	110	23 870
71·0	16·0	3·0	90	17 910	54·6	15·4	—	70	14 350
34·0	15·9	1·1	51	8 670	27·9	3·1	—	31	6 231
27·0	6·0	1·0	34	6 154	30·3	2·7	—	33	5 346
10·8	3·2	—	14	2 464	5·0	—	—	5	840
7·0	—	—	7	1 190	5·0	—	—	5	710

* Some of the groups have fractional numbers because one or two hairs were lost between weighing and classifying, and these were assumed to be distributed proportionately in all classes. The weights were supplied as mean weights per centimetre of hair to the nearest unit, and those in this table are the means multiplied by the number of hairs in the group. The effect of these approximations in the data is exceedingly small.

To find a, b and c we solve equations (11.6), and for these we need to determine the sums of squares and products from the data. It is not fair, however, to give all groups, whether large or small, the same weight in finding these sums. If the number of hairs in a group is n, and the variability of weights of single hairs is approximately the same in all classes, with a standard deviation of σ, the standard error of the mean weight per hair for the group is $\sigma/\sqrt{n}$, and that of the total weight is n times this, and so is proportional to $\sqrt{n}$; hence the variance of the total weight of a group due to variations between hairs within the same class is proportional to n, the number in the group. We may use the results of section 3.7 and regard $1/n$ as the quantity of information; then in making summations, each term may be weighted with this quantity, multiplying each product and square by $1/n$ before summing. For example, the weighted sum,

$$Syx_1 = \left(\frac{3\,243 \times 15{\cdot}4}{23} + \frac{5\,580 \times 25{\cdot}0}{30} + \ldots \right)$$

A further modification is necessary; a group with no hairs naturally has zero weight, so we shall make our regression line pass through the origin and not through the mean, and shall measure all deviations from zero; that is, in equations (11.5) and (11.6) we write x and y instead of x' and y'. It may assist some to think of the data of Table 11.8 as a half of the complete data, the other half having equal negative values, so that the mean is zero, and the sums of squares and products we obtain are half the total for the imaginary complete results.

The appropriate sums of squares and products are set out in equations below, and their solution yields the regression equation following them.

$$\begin{aligned} 363\,785{\cdot}7 &= 1\,468{\cdot}85a + 288{\cdot}83b + 18{\cdot}42c \\ 73\,472{\cdot}3 &= \quad 288{\cdot}83a + \quad 74{\cdot}73b + \quad 4{\cdot}94c \\ 4\,646{\cdot}0 &= \quad\ 18{\cdot}42a + \quad\ 4{\cdot}94b + \quad 1{\cdot}04c \end{aligned}$$

$$Y = 226{\cdot}255\,7x_1 + 114{\cdot}127x_2 - 82{\cdot}16x_3$$

Although the data are to some extent approximate, the arithmetic must be performed with considerable precision if the final results are to have any accuracy at all. In finding Sx_1y/n we calculated x_1/n to five decimal places, and then summed $y \times x_1/n$ on a machine;

as a check, we then calculated y/n and summed $x_1 \times y/n$. In performing the divisions for the solution of the simultaneous equations we had to work to nine or ten significant figures to obtain the accuracy of the constants shown in the regression. It is not claimed that the four decimal places for a have any physical meaning, but as a constant of the sample which will be used to find the sum of squares of group weights, that accuracy is necessary. Classes A and B have differing mean hair-weights per centimetre, while the hairs of C appear to have a negative one; however, there were very few of class C, and this irrational result is due to sampling variations, and signifies nothing. To find the sum of squares associated with the regression we use equation (11.7), and find

$$226{\cdot}255\,7 \times 363\,785{\cdot}7 + 114{\cdot}127 \times 73\,472{\cdot}3 - 82{\cdot}16 \times 4\,646{\cdot}0 = 90\,312\,000.$$

This could not have been given correct even to five figures if we had determined the regression coefficients correct only (say) to three. The weighted sum of squares of the hair-weights is

$$\frac{3\,243^2}{23} + \frac{5\,580^2}{30} + \ldots = 91\,220\,148,$$

and we will use this to five significant figures only.

We now have to find a regression equation using one constant, the mean hair-weight per centimetre for all classes. Using only the first term and equation of (11.6), and finding the weighted sums of products and squares, we have

$$S\left(\frac{yx_1}{n}\right) = a'S\left(\frac{x_1^2}{n}\right);$$

since there is only one class, $x_1 = n$, and we obtain $Sy = a'Sn$, which is the straightforward relationship, *mean hair-weight = total weight of all groups divided by the total number of hairs*. This gives $a' = 203{\cdot}736$, and the sum of squares of weights associated with it is $203{\cdot}736 \times 441\,904 = 90\,032\,000$ (correct to five significant figures).

The analysis of variance is in Table 11.9, and in reckoning the degrees of freedom it must be remembered that deviations are not measured from a mean determined from the data, but from the origin, so that the fifty groups contribute fifty degrees to the sum of

squares. In order to test the significance of the difference in regressions, we find

$$z = \tfrac{1}{2}\log_e \frac{140}{19{\cdot}3} = 0{\cdot}99,$$

and this is significant, for when $n_1 = 2$ and $n_2 = 30$, $z = 0{\cdot}84$ lies on the 1 per cent. level. Thus we conclude that the two classes of hairs *A* and *B*, as determined by swelling in caustic soda, have significantly different mean hair-weights per centimetre. This result could not have been obtained directly, because there were no means of weighing the hairs individually, and it was not practicable to obtain the individual classes for separate weighing.

TABLE 11.9

ANALYSIS OF VARIANCE OF WEIGHTS OF GROUPS OF COTTON HAIRS

Source of Variation	Sums of Squares		Degrees of Freedom		Variance
Multiple regression—					
Simple regression	90 032 000	90 312 000	1	3	90 032 000
Difference in regression ..	280 000		2		140 000
Residual	908 000		47		19 300
Total	91 220 000		50		—

11.8. We must emphasise the underlying assumption of the methods of this chapter that the residuals are homogeneous in the sense that their variance and co-variance are constant. In the example of section 11.1, for instance, we assume the relation between residual straw and grain yields is sensibly the same for all treatments and blocks, and in the example of section 11.5 illustrating multiple regression, we are assuming that the relation between loaf volume and protein extracted is the same for all values of crude protein. The tests of significance also assume the residual deviations to be normally distributed.

APPENDIX

Proof of the Formula for the Partial Correlation Coefficient

Let x and y be two variables, and let z be the third which is to be eliminated. Then the regressions of x and y on z (measuring all as deviations from the mean) are

$$X' = r_{xz}\sqrt{\frac{Sx'^2}{Sz'^2}} \cdot z'$$

and . . . . . (11.8)

$$Y' = r_{yz}\sqrt{\frac{Sy'^2}{Sz'^2}} \cdot z';$$

and the partial correlation coefficient between x and y is

$$r_{xy\,.\,z} = \frac{S(x' - X')(y' - Y')}{\sqrt{S(x' - X')^2 S(y' - Y')^2}} \quad . \quad . \quad . \quad (11.9)$$

Expanding the numerator of (11.9)

$$S(x' - X')(y' - Y') = Sx'y' + SX'Y' - Sx'Y' - SX'y',$$

and substituting from (11.8) for X' and Y',

$$\begin{aligned} S(x' - X')(y' - Y') &= Sx'y' + r_{xz}r_{yz}\sqrt{Sx'^2Sy'^2} - r_{yz}\sqrt{\frac{Sy'^2}{Sz'^2}}Sx'z' \\ &\quad - r_{xz}\sqrt{\frac{Sx'^2}{Sz'^2}}Sy'z', \\ &= Sx'y' - r_{xz}r_{yz}\sqrt{Sx'^2Sy'^2}, \\ &= \sqrt{Sx'^2Sy'^2}(r_{xy} - r_{xz}r_{yz}). \quad . \quad . \quad . \quad . \quad . \quad (11.91) \end{aligned}$$

Expanding the first term in the denominator of (11.9),

$$S(x' - X')^2 = Sx'^2 + SX'^2 - 2Sx'X',$$

REFERENCES

BALLS, W. L., and HANCOCK, H. A. (1926). Measurements of the reversing spiral in cotton hairs. *Proc. Roy. Soc., B*, xcix, 130.

BATESON, W. (1913). Mendel's Principles of Heredity. Cambridge University Press. 345.

BOWLEY, A. L. (1926). Elements of Statistics. 5th edition. King.

CHRISTIDIS, B. G. (1931). The importance of the shape of plots in field experimentation. *Jour. Agri. Sci.*, xxi, 14.

COCHRAN, W. G. (1936). Statistical analysis of field counts of diseased plants. *Supp. Jour. Roy. Stat. Soc.*, iii, 49.

COLLINS, G. N., FLINT, L. H., and MCLANE, J. W. (1929). Electric stimulation of plant growth. *Jour. Agri. Research*, xxxviii, 585.

CO-OPERATIVE STUDY (1917). On the distribution of the correlation coefficient in small samples. *Biometrika*, xi, 328.

— (1923). On the nest and eggs of the Common Tern. *Biometrika*, xv, 294.

CORKILL, B. (1930). The influence of insulin on the distribution of glycogen in normal animals. *Biochem. Jour.*, xxiv, 779.

DARBISHIRE, A. D. (1904). On the result of crossing Japanese Waltzing with Albino mice. *Biometrika*, iii, 1.

EDEN, T., and FISHER, R. A. (1927). The experimental determination of the value of top dressings with cereals. *Jour. Agri. Sci.*, xvii, 548.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, x, 507.

— (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, i, No. 4.

— (1922). On the mathematical foundations of statistics. *Phil. Trans. Roy. Soc.*, A, ccxxii, 309.

— (1922*b*). On the interpretation of χ^2 from contingency tables, and the calculation of *P. Jour. Roy. Stat. Soc.*, lxxxv, 87.

— (1924*a*). On a distribution yielding the error functions of several well known statistics. *Proc. of the International Math. Congress, Toronto*, 805.

— (1924*c*). The distribution of the partial correlation coefficient. *Metron*, iii, No. 3.

— (1925*a*). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, xxii, 700.

— (1925*b*). Applications of "Student's" distribution. *Metron*, v, No. 3.

— (1928). Triplet children in Great Britain and Ireland. *Proc. Roy. Soc.*, B, cii, 286.

— (1930*a*). The moments of the distribution for normal samples of measures of departure from normality. *Proc. Roy. Soc.*, A, cxxx, 16.

— (1930*b*). Inverse probability. *Proc. Camb. Phil. Soc.*, xxvi, 528.

— (1935). The logic of inductive inference. *Jour. Roy. Stat. Soc.*, xcviii, 39.

— (1936 or 1936*a*). Statistical Methods for Research Workers. 6th Edition. Oliver & Boyd.

— (1936*b*). The Design of Experiments. 2nd Edition. Oliver & Boyd.

FISHER, R. A., and HOBLYN, T. N. (1928). Maximum- and minimum-correlation tables in comparative climatology. *Geografiska Annaler*, iii, 267.

—, THORNTON, H. G., and MACKENZIE, W. A. (1922). The accuracy of the plating method of estimating the density of bacterial populations. *Annals of Appl. Biology*, ix, 325.

GLANVILLE, W. H., and REID, D. A. G. (1934). Mortar tests as a guide to the strength of concrete. *Structural Engineer*, xii, 242.

GOULD, C. E., and HAMPTON, W. M. (1936). Statistical methods applied to the manufacture of spectacle glasses. *Supplement Jour. Roy. Stat. Soc.*, iii, 137.

HARMON, G. E. (1926). On the degree of relationship between head measurements and reaction time to sight and sound. *Biometrika*, xviii, 207.

HARRIS, J. A. (1910). On the selective elimination occurring during the development of the fruits of *Staphylea*. *Biometrika*, vii, 452.

HARRIS, R. H. (1931). Relation of peptization of wheat flour protein to loaf volume. *Cereal Chemistry*, viii, 47.

IRWIN, J. O. (1931). Mathematical theorems involved in the analysis of variance. *Jour. Roy. Stat. Soc.*, xciv, 284.

JONES, H. G. (1910) (data for a note by K. P.). On the value of the teachers' opinion of the general intelligence of school children. *Biometrika*, vii, 542.

KOSHAL, R. S., and TURNER, A. J. (1930). Studies in the sampling of cotton for the determination of fibre-properties. *Jour. Textile Inst.*, xxi, T325.

LATTER, O. H. (1902). The egg of *Cuculus canorus*. *Biometrika*, i, 164.

MAHALANOBIS, P. C. (1932). Auxiliary tables for Fisher's z-test in analysis of variance. *Indian Jour. of Agri. Sci.*, ii, 679.

— (1933). Tables for the application of L-tests. *Indian Jour. of Statistics*, i, 109.

MATTHEWS, J. R. (1923). The distribution of certain portions of the British flora. *Annals of Botany*, xxxvii, 277.

MINER, J. (1922). Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for Use in Partial Correlation and Trigonometry. Johns Hopkins Press, Baltimore.

MORTON, W. E. (1926). The importance of hair weight per centimetre as a measurable character of cotton and some indications of its practical utility. *Jour. Textile Inst.*, xvii, T537.

MUMFORD, A. A., and YOUNG, M. (1923). The interrelationships of the physical measurements and the vital capacity. *Biometrika*, xv, 109.

NAYER, P. P. N. (1936). An investigation into the application of Neyman and Pearson's L_1 test, with tables of percentage limits. *Statistical Research Memoirs*, i, 38.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Jour. Roy. Stat. Soc.*, xcvii, 558.

— and PEARSON, E. S. (1931). On the problem of k samples. *Bull. de l'Acad. Polonaise des Sci. et des Let.*, Série A, 460.

— —, (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, xxA, 175.

ORENSTEEN, M. M. (1920). Correlation of cephalic measurements in Egyptian born natives. *Biometrika*, xiii, 17.

PARKES A. S., and DRUMMOND, J. C. (1925). The effect of vitamin B deficiency on reproduction. *Proc. Roy. Soc.*, B, xcviii, 147.

PEARL, R. (1907). Variation and Differentiation in *Ceratophyllum*. Carnegie Institution of Washington.

PEARSE, G. E. (1928). On corrections for the moment-coefficients of frequency distributions. *Biometrika*, xxA, 314.

PEARSON, E. S. (1926). A further note on the distribution of range in samples taken from a normal population. *Biometrika*, xviii, 173.

— (1930). A further development of tests for normality. *Biometrika*, xxii, 239.

— (1932). The percentage limits for the distribution of range in samples from a normal population. *Biometrika*, xxiv, 404.

PEARSON, K. (1895). Skew variation in homogeneous material. *Phil. Trans. Roy. Soc.*, A, clxxxvi, 343.

— (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.*, Series V, l, 157.

— (1904). On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Research Memoirs*, i.

— (1905). On the general theory of skew correlation and non-linear regression. *Drapers' Company Research Memoirs*, ii.

— (1910). On a new method of determining correlation. *Biometrika*, vii, 248.

— (1920). On the probable errors of frequency constants. *Biometrika*, xiii, 113.

— (1931) (Editor). Tables for Statisticians and Biometricians. Part I, 3rd Edition. Cambridge University Press.

— and LEE, A. (1903). On the laws of inheritance in man. *Biometrika*, ii, 357.

PRZYBOROWSKI, J., and WILEŃSKI, H. (1935). Statistical principles of routine work in testing clover seed for dodder. *Biometrika*, xxvii, 273.

SANDERS, H. G., and WISHART, J. (1935). Principles and Practice of Field Experimentation. Empire Cotton Growing Corporation.

SMITH, A. M., and PRENTICE, E. G. (1929). Investigation on *Heterodera schachtii* in Lancashire and Cheshire. *Annals of Applied Biology*, xvi, 324.

SNOW, E. C. (1911). On the determination of the chief correlations between collaterals in the case of a simple Mendelian population mating at random. *Proc. Roy. Soc.*, B, lxxxiii, 37.

"STUDENT" (1907). On the error of counting with a hæmacytometer. *Biometrika*, v, 351.

— (1908). The probable error of a mean. *Biometrika*, vi, 1.

— (1925). New tables for testing the significance of observations. *Metron*, v, No. 3, 18.

TIPPETT, L. H. C. (1927). Random sampling numbers. *Tracts for Computers*, xv. Cambridge University Press.

— (1934). Statistical methods in textile research; uses of the binomial and Poisson distributions. *Shirley Inst. Mem.*, xiii, 35, or *Jour. Text. Inst.*, 1935, xxvi, T13.

TOWER, W. L. (1902). Variation in the ray-flowers of *Chrysanthemum leucanthemum*. *Biometrika*, i, 309.

TSCHEPOURKOWSKY, E. (1905). Contributions to the study of interracial correlation. *Biometrika*, iv, 286.

WARREN, E. (1909). Some statistical observations on Termites. *Biometrika*, vi, 329.

WELDON, W. F. R. (1901). Change in organic correlation of *Ficaria ranunculoides* during the flowering season. *Biometrika*, i, 125.

WHELDALE, M. (1907). The inheritance of flower colour in *Antirrhinum majus*. *Proc. Roy. Soc.*, B, lxxix, 288.

WINTER, F. L. (1929). The mean and variability as affected by continuous selection for composition in corn. *Jour. Agri. Research*, xxxix, 451.

WISHART, J. (1928). Table of significant values of the multiple correlation coefficient. *Quarterly Jour. Roy. Met. Soc.*, liv, 258.

YATES F. (1933*a*). The formation of Latin squares for use in field experiments. *Empire Jour. Exp. Agri.*, i, 235.

(1933*b*). The principles of orthogonality and confounding in replicated experiments. *Jour. Agri. Sci.*, xxiii, 108.

(1933*c*). The analysis of replicated experiments when the field results are incomplete. *Empire Jour. Exp. Agri.*, i, 129.

(1936*a*). Incomplete Latin squares. *Jour. Agri. Sci.*, xxvi, 301.

— (1936*b*). A new method of arranging variety trials involving a large number of experiments. *Jour. Agri. Sci.*, xxvi, 424.

— (1936*c*). Incomplete randomised blocks. *Annals of Eugenics*, vii, 121.

YULE, G. U. (1927). An Introduction to the Theory of Statistics, 7th Edn. Griffin.

ZINN, J. (1923). Correlations between various characters of wheat and flour as determined from published data, etc. *Jour. Agri. Research*, xxiii, 529.

INDEX

ages of husbands and wives, 161
American bladder nut, 126
analysis of variance, 126, 213
— —, computation, 130, 217, 226, 232
— —, non-uniform arrays, 132
— —, relation to correlation, 156
analysis of co-variance, use for increasing precision of experiment, 248
arbitrary origin and scale, 36
arithmetic mean, 28
array, 126
array means, curve of, 154, 192
association, 106, 126, 140
—, measurement of, 129, 160, 190, 202
asymmetry, 34
attribute, 16
average, 28

bacterial counts, 123
Balls, 20
Barlow, 18
Bateson, 103
Beavan, 235
β, 34
—, sampling errors, 86
binomial distribution, 46
— —, moments, 48
— —, small samples, 121
— —, standard error of mean, 82
binomial distribution, use for testing randomness, 50
Bowley, 209
British Association, 73
Brownlee, 106

cabbage seed, 50, 102
causation, 162
cephalic index, 178
Ceratophyllium, 22
chance, 43, 44
character, 16
chess-board arrangement, 229
χ^2, 98
χ^2, additive nature, 102
—, general character, 108
—, sampling distribution, 98, 108, 271
—, use for testing significances of correlations, 178
choice of statistical constants, 92
Christidis, 228
chrysanthemum, 23
cloudiness, 23
Cochran, 54
coefficient of variation, 30
— —, standard error, 88
Collins, 114
complex samples, standard errors, 204
computation, *see under various constants*
—, accuracy in, 42, 266
concrete, 182
confidence limits and coefficient, 90
confounding, 241
consistency, 92
constant, statistical, 19
contingency, 105
—, coefficient, 202
—, relation to correlation, 203
—, table, 104
continuous variate, 16
control, 15
controls, use of, 234
Co-operative study, 142, 176
Corkill, 115
corn yield, 79, 227
correlation, 140
—, analysis of variance, 156, 192, 258
—, multiple, 257
—, partial, 243
—, partial, linear, 251
—, partial, non-linear, 253
—, spurious, in time series, 254
correlation coefficient, 144
— —, combination of estimates, 177
— —, computation, 165
— —, estimation, 164

correlation coefficient, multiple, 257
— —, multiple, sampling errors, 262
— —, partial, 251
— —, partial, proof of formula, 268
— —, partial, sampling errors and significances, 261
— —, significance of differences, 176, 179
— —, standard error, 172
— —, tests of significance, 173
— —, transformation, 176
correlation ratio, 189
correlation table, 140
cotton hairs, lengths, 70
— —, sampling, 68, 70, 205
— —, spiral reversals, 20
— —, weights, 263
cotton seed, 122
co-variance, 246
—, use to increase precision of experiment, 248
cuckoos' eggs, 133, 191
cumulative diagram, 20
curve fitting, frequency, 50, 56
— —, regression, 149, 195, 256
cyst counts, 123, 207

Darbishire, 83
degrees of freedom, 99, 110
— —, effect of fitting constants, 99, 159
difference between two samples, standard error, 72
—, significance, 98, 108
discrete variate, 16
dispersion, 30
distribution, 19
Drummond, 83

economy in sampling, 207
Eden, 243
efficiency, 92
Egyptians, 180
Elderton, 62, 100
electric current, effect on growth, 113
electric lamps, 30
elementary schoolboys, 161
Englishmen, heights, 73, 84, 118
error curve, 54
errors, 17
—, of first and second kinds, 74
estimation, theory of, 92
event, 43, 47
experimental arrangement, 210
experimental method, 15
experiments, agricultural, 226

factorial experiments, 235
fathers' heights, 39, 101
fiducial limits and probability, 90
Fisher, 18, 54, 57, 86, 90, 92, 96, 100, 109, 113, 117, 123, 147, 159, 162, 173, 176, 177, 197, 230, 240, 241, 243, 247, 250, 261
flatness, of mode, 33
Flint, 114
flower colour, 102
Fourier analysis, 50
fraternal correlations, 177
frequency constants, 27
frequency curves, 54, 60
— —, non-normal, 61
— —, normal, 54
frequency diagram, 20
frequency distributions, 19
— —, comparison between experimental and theoretical, 98
— —, comparison between experimental, 107
frequency polygon, 20
frequency surface, 144
frequency table, 19
— —, formation, 26
functions of statistical constants, standard errors, 87

Gaussian distribution, 54
geometric mean, 28
Glanville, 182
glycogen, 116, 119
goodness of fit, 98
Gould, 218
grain yields, 227
grain and straw yields, 243

grouping, 26, 140
groups of correlation coefficients, 178
groups of samples, significance of means, 78, 136

haemacytometer, 51
half-drill strip, 235
Hampton, 218
Hancock, 20
Harmon, 148
harmonic mean, 28
Harris, J. A., 126
Harris, R. H., 257
head length, 148, 180
heterogeneity of variability, 138
— —, effect on theory of errors, 204
— —, effect on correlation, 243
histogram, 20
Hoblyn, 147, 162, 247
hollow curve, 26
houses, sampling, 68

independence, 44
individuals, 16
inductive inference, 89
infinite population, 45, 62, 67
insulin, 115, 119
insurance, 17
interaction, 222, 236
interpolation, in table of z, 119
—, linear, 52, 57
intrinsic accuracy, 93
inverse probability, 91
Irwin, 64, 159

Jones, 161

kinetic theory of gases, 27
Koshal, 70
kurtosis, 33

Latin square, 230
— —, computation, 232
— —, formation, 233
Latter, 134
law of small numbers, 49
least squares, 155
— —, proof of linear equations, 167
leaves per whorl, 22
Lee, 39
level of significance, 70
— — and probability, distinction, 76, 136
likelihood, 95
loaf volume, 192, 257

Mackenzie, 123
Mahalanobis, 118, 120
maize seedlings, 113
marriages in churches, 162
mathematical probability, 44
Matthews, 23
maximum likelihood, 95
McLane, 114
mean, 28
—, differences expressed as variance, 128
—, sampling distribution, 63
—, significance, 70, 71
—, —, parallel pairs, 211
—, —, small samples, 112, 114
—, standard error, 62, 63
—, standard error in complex population, 204
mean deviation, 30
— —, standard error, 85
mean square contingency, 202
median, 29
Mendel, 103
mice, 83
Miner, 251
mode, 26, 29
moments, computation, 34
—, deduction from continuous frequency curve, 55
—, second, 30
—, third and fourth, 33
mortar, 182
Morton, 141
multiple factor experiment, 235
Mumford, 251

Nayer, 120
Neyman, 74, 90, 120

non-normal frequency curves, 61
normal distribution, 54, 270
— —, determination of frequencies, 56
— —, moments, 55
— —, practical applicability, 60
— —, relation between frequencies and standard deviation, 59
normal surface, 149
normality, tests for, 86

occurrence, 47
odds, *see* probability
ogive, 20
Orensteen, 180
orthogonal equations, 197
orthogonality, 241
ovules per ovary, 126, 191

pairs, variance from, 111
parameters, 92
parents and children, 161
Parkes, 83
peakiness, *see* kurtosis
Pearl, 22
Pearse, 23
Pearson, E. S., 74, 79, 85, 86, 120, 138
Pearson, K., 18, 31, 33, 39, 57, 86, 98, 100, 106, 176, 190
peas, 103, 121
Peter's method, 30
physical determinations, random errors, 69
pistils, 145, 177
Poisson distribution, 48
— —, small samples, 123
— —, standard error of mean, 84
— —, tables, 49
— —, testing randomness, 51
polynomial equation, 195
— —, system of fitting, 197
population, 16
—, determination from sample, 89
—, infinite, 45, 62, 67
Prentice, 123, 207
probability, 43
—, test of laws, 49
probability integral, 56
probable error, 71
product moment, 165
— —, computation, 166
proportionate frequencies, 28
protein content, of flour, 257
— —, of wheat, 174, 192, 197, 254
Przyborowski, 54

qualitative and quantitative variate, 16
quantity of information, 94, 178, 265
quartile deviation, 30
— —, relation to standard deviation, 60

rabbits, 115, 119
rainfall and sunshine, 161
random groups or blocks, 230
random samples, 45, 67
— —, complex theory, 204
— —, technique, 67
random sampling numbers, 68
randomness, tests of, 49
range, 31
—, standard error, 85
ratio between two means, standard error, 88
reaction time, 148
recruits, 20
regression, 149
—, as variance, 193, 200
—, line through origin, 265
—, linear, 149
—, multiple, 256
—, non-linear, 189
—, partial, 256
—, tests for linearity, 192, 200
regression coefficients, 155
— —, computation, 165
— —, estimation, 164
— —, significances, 181
— —, standard error, 182
Reid, 182
representative samples, 67
residuals, 130
—, correlation of, 244, 253

routine analysis, determination of errors, 206

sample, 16
—, random, 45, 67
—, small, 110
sampling, by strata, 209
—, unrestricted random method, 209
sampling distribution, 62
— —, skew, 80
— —, of various constants, *see under the constants*
sampling from limited field, 209
sampling technique, 68
sampling theory, complex populations, 205
— —, small samples, 110
Sanders, 240, 250
scatter, *see* dispersion
scatter diagram, 140
Scotsmen, heights, 73, 84, 118
semi-interquartile distance, 31
set, 47
sex-ratio of rats, 83, 104
shape of distribution, 33
Sheppard, 57
Sheppard's corrections, 39, 131
— —, effect on tests for correlation, 188
significance, statistical, 71
—, from skew distribution, 80
—, tests of, 69
—, *see also under various constants*
skew sampling distributions, tests of significance, 80
skewness, 21, 34
—, test for, 86
small numbers, law of, 49
smallpox, 106, 203
small samples, 110
Smith, 123, 207
smoothing, 201
Snow, 161, 177
Soper, 49
spectacle glass, 218
spiral reversal, 20
spurious correlation, 254
stamens, 145, 177
standard deviation, 30
— —, estimated from two samples, 73
— —, relation to normal frequencies, 59
— —, standard error, 84
standard error, 62
— —, complex samples, 204
— —, of various constants, *see under the constants*
statistical method, 15
statistical probability, 44
statistical significance, *see* significance
statistics, 92
strata, 209
Student, 52, 113
sub-range, 19
summation, rules, 37
sunshine and rainfall, 161
survivor curve, 20

t test, 112, 271
—, essential character, 114
—, relation to *z* test, 134
temperatures, maximum and minimum, 147, 247
termites, 137, 211
tern eggs, 142, 157
Thornton, 123
time series, 197, 201
— —, correlation of, 254
Tippett, 54, 68, 120
Tower, 23
transformed variate, 36
trial, 47
triplet births, 177
Tschepourkowsky, 178
Turner, 70

universe, 16

vaccination, 106, 203
variability, 30
variable, 16
—, dependent and independent, 154
variance, 30
—, additive nature, 125

variance, analysis, 125
—, analysis, computation, 130, 134, 217, 232
—, analysis, effect of skewness, 138
—, analysis into many parts, 213
—, associated with regression, 156, 192, 200, 259
—, computation, 34
—, estimation from small samples, 110
—, sampling errors in small samples, 117, 120
—, standard error, 88
variate, 16
vice-counties, 23
vital capacity, 251
vitamin B, 83, 105

waltzing mice, 83
Warren, 137
weaving of cotton warps, 235
weighting, *see* quantity of information
Weldon, 145, 161
Wheldale, 102
Wileński, 54
Willis, 26
Winter, 175
Wishart, 240, 250, 262
working mean, 36

Yates, 18, 234, 241
yeast cells, 51, 84
Young, 251
Yule, 161, 162

z, interpolation in tables of, 119
z test, 117, 272
—, relation to t test, 134
z', transformation for correlation coefficient, 176
Zinn, 192